BAYESIAN MODELLING OF A HAPTIC CATEGORIZATION TASK, AND TEMPORAL VIDEO SEGMENTATION

BY
NOAH RIPSTEIN

A Thesis
Submitted to Department of Psychology, Neuroscience and Behaviour
In Partial Fulfilment of the Requirements
for the degree
Bachelor of Arts and Science with Combined Honours in Psychology, Neuroscience and Behaviour

McMaster University
April 2024

HONOURS BACHELOR OF ARTS AND SCIENCE

MCMASTER UNIVERSITY

Hamilton, Ontario

TITLE: Bayesian Modelling of a Haptic Categorization Task, and Temporal Video Segmentation

AUTHOR: Noah Ripstein

SUPERVISOR: Dr. Daniel Goldreich

NUMBER OF PAGES: v, 44

Abstract

As we interact with an object using our hands, the human sensory system provides the brain with uncertain information which is used to inform an inference about the nature of the object. The human sensory experience is shaped by these inferences, which can be formalized with probabilistic models. Previous research, which focused primarily on modelling human predictions about continuous variables, has indicated that many aspects of human sensory perception are well-approximated by Bayesian inference. We devised a haptic categorization task in which participants repeatedly classify ambiguous novel stimuli into one of two probabilistic categories which are defined by two-dimensional Gaussian distributions according to independent tactile features. After each trial, participants were informed whether their classification was correct. We compared human performance to that of various Bayesian observers which mathematically describe various classification strategies. We found that most participants used one of two strategies: optimally integrating both independent features, and performing optimal inference using only the tactile feature which elicits the least sensory noise. Our findings support the hypothesis that human sensory perception is guided by Bayesian inference for categorization tasks.

Videos of participants performing the experiment were collected, and we designed a machine learning computer vision system which can automatically detect the duration of each trial. Our system is trained only on a broad public dataset, which avoids the need for time-consuming, task-specific data labeling. Our system produces predictions with promising preliminary results.

Acknowledgements

I would like to extend my sincere gratitude to everyone who supported me throughout the journey of completing this thesis. It would not have been possible without the collective guidance I received. First and foremost, I am profoundly grateful to my thesis advisor, Dr. Daniel Goldreich, for his invaluable mentorship, encouragement, and expertise. I am especially thankful for his willingness to support and learn with me as I decided to add a large computer vision component to my thesis. My appreciation also extends to other members of the lab, including Yao and Lyljana, and to Grace, Steven, and Shafee, whose help and support were invaluable. I would also like to thank my parents, my friends, and especially Maya Mammon for their enthusiastic support and encouragement.

Contents

List of Tables

List of Figures

Introduction

We can model human perception as a series of probabilistic inferences: the sensory system provides the brain with uncertain information about the environment, and we must use this imperfect information to form an understanding of the world around us to form a perception. When we hear our friend call out to us, our sensory system converts their voice to neural impulses, which our brain uses to infer who and where they are before we turn to see them. When we then turn around to see a familiar figure waving from across the street, we update our beliefs by incorporating the new data that is presented to the visual cortex. In doing so, we could either become more confident in our hypotheses about the identity and location of the caller, or use this additional sensory data to infer that our initial conviction is mistaken. The human sensory experience is shaped by these inferences, not just by the neural signal produced. Such probabilistic reasoning can be modeled using Bayesian inference, which incorporates current sensory data and previous experiences to produce a posterior probability distribution for the caller's identity and a posterior probability density for their location. Such a Bayesian framing has prompted significant scientific inquiry into statistical modelling of perception, sparking a paradigm shift towards understanding perception as an unconscious exercise in Bayesian inference.

Formally, this Bayesian perceptual framework holds that every sensory perception is conceptualized as a posterior probability, $P(H|D)$. This posterior probability represents our revised belief in a hypothesis H after taking into account new data D. It is calculated using Bayes' theorem (Equation 1.1), which combines two key components: the likelihood and the prior. The likelihood, $P(D|H)$, is the probability of observing the data assuming that our hypothesis is true. It reflects how well the data supports the hypothesis. The prior, $P(H)$, represents our initial belief in the hypothesis before considering the new data: it encapsulates our previous knowledge or assumptions about the hypothesis. Bayes' Theorem is often written ignoring the normalization constant in the denominator: the posterior is proportional to the product of the likelihood and prior (Equation 1.1).

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \quad \text{equivalently,} \quad P(H|D) \propto P(D|H)P(H) \tag{1.1}$$

Much of the research in perception as Bayesian inference has focused on comparing human participants to a "Bayes-optimal observer" (O'Reilly et al., 2012). Such Bayesian observers are computational models which complete a task as similar as possible to that of human participants,

always relying on an underlying mathematical strategy to determine perception. A Bayes-optimal observer is one which uses Bayes' Theorem to weigh the reliability of cues in order to determine optimally the characteristics of stimuli with imperfect information (Alais and Burr, 2004). Designing observers in this way allows researchers to determine how human participants would perform if their perceptual inference followed a given model.

There are several perceptual processes in humans which can be well-approximated by a Bayes-optimal observer (Shams and Beierholm, 2022; Colombo and Seriès, 2012). Audio-visual tasks in particular have been closely modelled. In addition to accurately modelling perception and categorization of such stimuli individually, these cues appear to be combined in a manner similar to a Bayes-optimal observer when presented simultaneously (Bankieris et al., 2017; Alais and Burr, 2004; Bejjanki et al., 2011).

Alais and Burr (2004) aimed to determine whether the ventriloquist effect—the perceptual phenomenon where a sound is perceived to originate from the location of a visual stimulus rather than its actual source—arises from a nearly statistically optimal combination of audio and visual cues, as weighted by uncertainty. Participants underwent a series of trials in which they were presented with a visual flash and auditory click separated by 500ms, and were asked to determine which stimulus was presented further to the left. The researchers manipulated the visual information that participants could obtain by adjusting the blur of the flash, and adjusted apparent sound-cue location by increasing interaural time differences. They found that audio-visual cues are combined near-optimally, where higher weight is assigned to cues with less uncertainty. The ventriloquist effect is not explained by vision "overriding" audio cues; rather, it is explained by vision tending to yield a lower degree of uncertainty than audio, meaning that people tend to assign more weight to vision, who thereby conclude that the audio cue originated from the same location as the visual cue. This optimal weighting of cues according to uncertainty provides credibility to the Bayesian model of sensory perception in cue combination.

Gepshtein and Banks (2003) sought to determine how visual and haptic information are integrated, and to mathematically model the degree and the circumstances by which independent sensory cues interact to inform perception. The first experiment in their study was a within-modality experiment, where participants were asked to compare two stimuli presented either visually or haptically, and judge which stimulus had a larger intersurface distance. By measuring participants' performance at different viewing angles, the researchers determined the just-noticeable difference for each modality and viewing condition. They found that visual performance was best when the surfaces were unobscured, whereas haptic performance remained

relatively constant across all viewing conditions.

In the second experiment conducted by Gepshtein and Banks (2003), participants used vision and touch simultaneously: they were presented with pairs of stimuli with conflicting visual and haptic information. Participants were again asked to judge which stimulus had a larger intersurface distance, this time using both touch and vision. In this series of trials, the researchers manipulated the viewing angle to examine how the participants combine the two sensory modalities under different degrees of visual uncertainty. They found that when combining cues, if the visual field is obstructed, then touch plays a stronger role in perceptual inference: perceptual cues were weighted according to their reliability.

Bankieris et al. (2017), who had participants categorize objects using sensory cue combination, found that participants combine audio-visual cues in a statistically optimal manner. They taught participants novel audio-visual categories with exemplars, rather than having participants classify stimuli into known categories, and had them group audio, visual, and audiovisual stimuli into these categories. Bankieris et al. (2017) devised a model that considers the degree of variance within audio-visual categories. This category variance causes some stimuli to be more exemplary of a given category than other stimuli. Their model acknowledges that both category variance and sensory variance—which arises from the limitations of human perception—affect performance on the task. Their model classifies audio, visual, and audio-visual stimuli statistically optimally into these categories by weighing cues based on the reliability associated with the two types of variance. They found that cue combination is statistically optimal, and that this leads to optimal categorization.

The focus of the present study is to determine the extent to which human performance aligns with a statistically optimal Bayesian observer's performance in a haptic categorization task. Identification of common objects using only touch is fast and accurate (Klatzky et al., 1985), and we construct a Bayesian model of this classification which we compare with human performance. The participants' task requires that they use only their sense of touch to classify novel objects into one of two novel categories. Each object has two independent tactile features, ensuring that the task is one of cue-combination and categorization. We will explore three hypotheses: that tactile perception on the task is guided by optimal Bayesian inference, or by sub-optimal Bayesian inference, or not guided by Bayesian inference in any form. In the context of this study, we define sub-optimal Bayesian inference to be the optimal inference using some, but not all, available sensory data. Inference guided by a Bayesian sub-optimal strategy results in participants making more classification errors. Participants use cue combination to form internal representations of the

novel categories, and use these representations to classify objects.

Our investigation constitutes a meaningful contribution to the field. Much of the foundational literature on sensory cue-combination as Bayesian inference has focused on using Bayesian modelling to estimate the posterior distribution of a continuous variable such as distance (Gepshtein and Banks, 2003) or location (Alais and Burr, 2004). Our model, on the other hand, must classify objects into one of two distinct categories. The inter-modality nature of our task further differentiates our investigation from other cue-combination literature: where Gepshtein and Banks (2003), Alais and Burr (2004) and Bankieris et al. (2017) combined cues from different sensory modalities, our task involves combining independent tactile cues.

Temporal Video Segmentation

Part of our experimental procedure included recording videos of participants completing trials repeatedly. Beyond Bayesian modelling, which is the primary focus of this work, we are interested in how long participants spend performing each trial. Such information is informative about the nature of participant classification strategies. Creating a machine learning algorithm which can automatically detect the duration of every trial in a video became a significant area of inquiry. Our method drew primarily on two areas of computer vision research: human-object interaction and Temporal Action Segmentation. Human-object interaction research focuses on identifying how someone is in contact with an object, usually one image at a time; Temporal Action Segmentation (TAS) is a type of machine video understanding which entails segmenting a single video into distinct temporal segments (Ding et al., 2024).

Primer on Object Detection

Object detection is a computer vision technique that identifies and locates multiple objects within an image or video by categorizing them and outlining their boundaries with bounding boxes (Amjoud and Amrouch, 2023). Unlike image classification, which categorizes the label of an entire image, object detection identifies multiple objects in an image, classifies them into different categories, and determines their precise locations with bounding boxes.

Modern object detection algorithms are primarily categorized into two types: two-stage detectors and one-stage detectors. Two-stage detectors, such as R-CNN (Uijlings et al., 2013) and Faster R-CNN (Ren et al., 2015) operate in a sequential manner. Initially, they generate region proposals—specific areas in the image likely to contain objects. These proposals are then classified

into object categories and their bounding boxes are refined. This process allows two-stage detectors to achieve high accuracy and precision, making them well-suited for applications where detection quality is more important than detection speed.

In contrast, one-stage detectors like YOLO (You Only Look Once (Redmon et al., 2016)) and SSD (Single Shot MultiBox Detector (Liu et al., 2016)) streamline this process by eliminating the region proposal stage. They directly predict object categories and bounding box coordinates across the entire image in a single pass through the network. This approach significantly speeds up the detection process, enabling real-time performance at the cost of accuracy. For our purposes, two-stage detectors are appropriate, because we analyze data after the experiment is completed, and therefore have no need to sacrifice accuracy for the sake of faster processing time.

Methods

2.1 Participants and Procedure

Twenty four undergraduate students (average age 19; 22 Female, 2 Male) at McMaster University served as our participants. All participants signed a consent form which states that they do not have any of the following: diabetes, nervous system disorder or injury, learning disability, dyslexia, attention-deficit disorder, cognitive impairment, carpal tunnel syndrome, arthritis of the hands, hyperhidrosis.

Participants completed a haptic categorization task, and we compared their performance to computer-simulated mathematical models which behave in accordance with Bayesian inference to complete the same task virtually. A Bayes-optimal model was simulated, along with several Bayesian sub-optimal strategies. The categorization task involves feeling a stimulus and attempting to accurately classify it into one of two novel categories labelled "Elyk" and "Noek".

Stimuli

We obtained a set of haptic stimuli roughly the shape of a large coin with dots on one face; stimuli differ by their number of sides, and by the distance between dots (Fig. 2.1A). There are 25 unique stimuli: every combination of 5 different dot spacings (4-8mm in 1 mm increments) and 5 different number of sides (6-10 sides). Each object appeared somewhere within the "Elyk" and "Noek" 2-dimensional Gaussian distributions, as determined by the object's number of sides and distance of dot spacing relative to the cateogry's mean and standard deviation. No object is perfectly described by either "Elyk" or "Noek". These categories are probability density functions (PDFs) centered at a mean number of sides and dot spacing; each object falls somewhere under the PDF for both categories (Equation 2.1). Objects near a category mean are most likely to be described by that category, but any object could be classified into either category (Fig. 2.1B).

Participants were tested in one of four experimental conditions (Table 2.1). Each condition differed in the standard deviations (category σ) for Elyk and Noek. When category σ values are higher, the width of the category distributions increases, thereby increasing the amount of overlap between the two categories. In all experimental conditions, the Gaussian distributions defining the categories have a spherical covariance structure (Equation 2.3). In all conditions, μ_E and μ_N are kept constant. Bayesian observers underwent the same experimental conditions as human

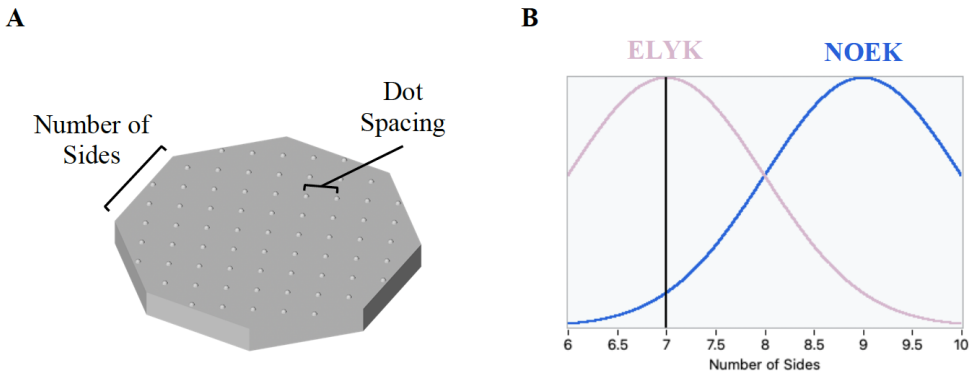


Figure 2.1: (A) Sample stimulus with 8 sides and 6 mm dot spacing. (B) 1D Gaussian distributions that define the sides feature of the Elyk and Noek categories. Elyk is centred at 7 sides with a standard deviation of 1, while Noek is centred at 9 sides with a standard deviation of 1. An object with 7 sides, for example, has a high probability of belonging to Elyk and a low probability of belonging to Noek.

participants (Table 2.1).

$$\text{Elyk} \sim \mathcal{N}(\mu_E, \Sigma_E) \quad \text{and} \quad \text{Noek} \sim \mathcal{N}(\mu_N, \Sigma_N) \tag{2.1}$$

$$\mu_E = \begin{bmatrix} 7 \\ 5 \end{bmatrix} \begin{matrix} \text{sides} \\ \text{dot spacing (mm)} \end{matrix} \quad \text{and} \quad \mu_N = \begin{bmatrix} 9 \\ 7 \end{bmatrix} \begin{matrix} \text{sides} \\ \text{dot spacing (mm)} \end{matrix} \tag{2.2}$$

$$\Sigma_E = \begin{bmatrix} \sigma_E & 0 \\ 0 & \sigma_E \end{bmatrix} \quad \text{and} \quad \Sigma_N = \begin{bmatrix} \sigma_N & 0 \\ 0 & \sigma_N \end{bmatrix} \tag{2.3}$$

Table 2.1: Experimental Conditions

Condition	σ_E	σ_N
Condition 1	0.75	0.75
Condition 2	0.75	1.25
Condition 3	1.25	0.75
Condition 4	1.25	1.25

Procedure

Before beginning the experiment, participants were informed that they will be given an equal number of “Elyk” and “Noek” objects. This serves to ensure that each participant’s prior probability for the category of an object is uniform. The experiment was conducted on two separate days, 7 days apart. Participants performed the procedure described below 405 times on each day over the course of roughly 2 hours (Fig. 2.2):

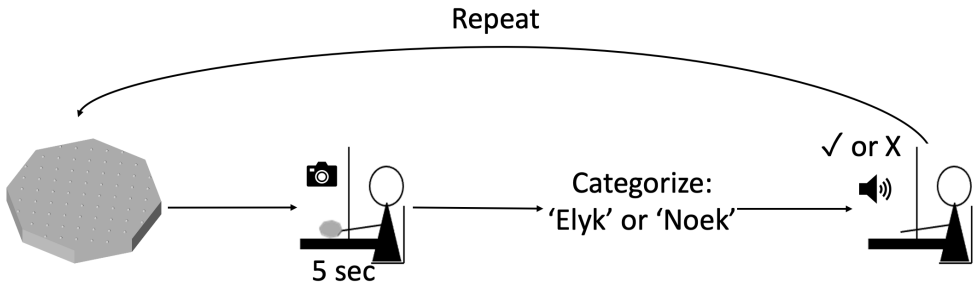
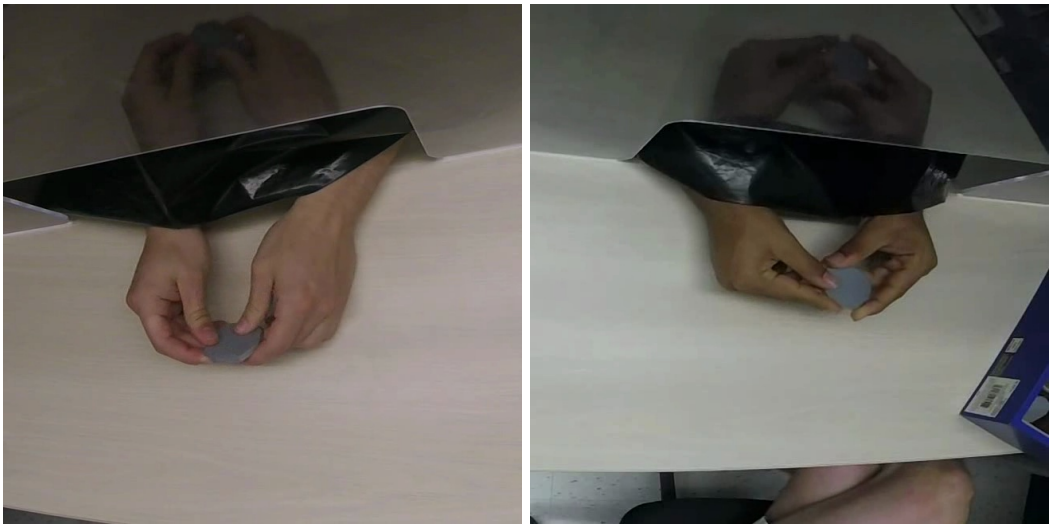


Figure 2.2: The procedure participants repeated 405 times on each of the two experiment days, which took place 7 days apart. An object was placed in front of the participant; they felt it for up to 5 seconds; they indicated whether they think the object belongs to the "Elyk" category or the "Noek" category; they received auditory indication about whether their categorization was correct.

1. The experimenter places an object on the table in front of the participant.
2. The participant feels the object with their hands for 5 seconds until they hear a beep, at which point they put it down.
3. The participant reports whether they believe the object is "Elyk" or "Noek."
4. The participant is informed whether they categorized the object correctly.

The trials were split into blocks of 45 consecutive trials. After blocks 3 and 6, participants were given a 5-minute break, and after all other blocks, they were given a 1-minute break. During each trial, the participant was seated behind an opaque screen with a cut-out for their arms so that they were unable to see the objects they were categorizing (Fig. 2.3). Objects were randomly selected by a computer program according to the following rules: during each trial, there is an equal probability of an "Elyk" and "Noek"; once the category is selected, the object given to the participant is selected by randomly sampling from the category's Gaussian distribution. Objects closer to the category means were thus selected more often throughout the experiment. After the participant identified to which category they believe the object belongs, an audio recording informed them whether they correctly categorized the object. This allowed them to learn the two categories using feedback provided throughout the experiment.

In every 45-trial block, overhead videos of 5 consecutive trials were taken of the participant's hands starting at the first trial of the block, the 20th trial, and the 41st trial. Sample screenshots of such videos are presented in Figure 2.3. Videos were collected for exploratory analysis so that we could identify relationships between performance and visual features in classification strategy. Identifying relationships between the amount of time a participant spends touching the object and their classification accuracy or the object's category probabilities could be informative and lead to



(a) Original camera angle (b) Modified camera angle

Figure 2.3: Screenshots of frames captured in recording of participant hands during classification. The apparatus which prevents the participants from visually inspecting the objects is visible. (A) and (B) are zoomed using the automatic zoom described in section 2.3. The visual differences between (A) and (B) are caused by the changes in camera mount, GoPro model and settings, which are adjusted to overcome technical difficulties.

future avenues of inquiry. Developing an automated method to detect this duration of each trial became a significant area of research, as is described in section 2.3.

2.2 Bayesian Modelling

We modelled Bayesian observers who simulated human participants on the haptic categorization task. Each Bayesian observer behaves according to a variation in Bayes-guided strategy. There are six observers in total (Table 2.3): one optimally combines sides and dots cues; another attends only to the number of sides; another attends only the density of dots; another attends to both features, but makes their inference by considering only the number of sides; another attends to both features, but makes their inference by considering only the number of dots. The final observer we consider is a null observer, which always guesses the category of the object randomly. The observers which discard part of the information optimally learn from the remaining information. The difference between an observer which attends to both sides and dots, but only makes an inference based on the number of dots and an observer which only observes the number of dots is in their sensory imprecision. This is discussed in more detail in the "Bayesian Observer Object Categorization" section.

Like human participants, Bayesian observers are never explicitly informed of the parameters

Table 2.2: Summary of Symbols used in Bayesian Modelling

Symbol	Description
μ_s	Mean number of sides for objects within a category
μ_d	Mean distance between dots for objects within a category
σ_s	Standard deviation in number of sides for objects within a category
σ_d	Standard deviation in distance between dots for objects within a category
D	Data collected on all trials
D_i	Data collected on trial i
$D_{<i}$	Data collected on all trials before trial i
ψ_s	Sensory imprecision of sides feature (variance of Gaussian distribution)
ψ_d	Sensory imprecision of dots feature (variance of Gaussian distribution)
S_s	Sensory measurement of sides feature
S_d	Sensory measurement of dots feature
ρ_g	The given piece on a trial (unknown to Bayesian observer)
ρ_j	A hypothesized piece given on a trial
H	Parameter vector containing σ_s, σ_d, μ_s, μ_d
"Elyk"	Participant classifies the object as Elyk
Elyk	Object truly comes from the Elyk distribution
m_k	Hypothesized measurement of piece k by participant
ω_s	True number of sides of a piece
ω_d	True dot spacing of a piece
$\mathcal{M}_j$	Participant is using model j as their classification strategy

which define "Elyk" and "Noek"; rather, using their designated learning strategy, they must form an internal approximation of the category parameters through repeated learning trials. Bayesian observers do make assumptions which human participants may or may not make. The observers assume that "Elyk" and "Noek" categories are defined by a 2D Gaussian distribution with a diagonal covariance structure: they attempt only to determine the distribution's mean and covariance matrix, not the type of distribution or the full nature of the distribution's covariance matrix structure. Equation 2.4 highlights the fact that the spherical covariance structure, which is the true structure of the category distribution covariance matrices, is a special case of the diagonal covariance matrix structure with $\sigma_1 = \sigma_2$. In both cases, the off-diagonal entries in the matrix (correlation parameters) are assumed to be 0.

Since Bayesian observers know that Gaussian distributions define the categories, observers also inherently assume that a category can be defined by a non-integer number of sides. Human participants might assume the opposite: they know that it is impossible for an object to have a non-integer number of sides; however, they might believe that the mean number of sides in a category is not an integer. They could reasonably expect, for example, that the mean number of sides in the "Elyk" category is 6.5. Additionally, rather than counting the number of sides directly, participants may use cues which are correlated with the number of sides as a proxy estimate for this

feature: side length and vertex sharpness are perfectly correlated with the number of sides. This enables participants to use continuous features to estimate the discrete number of sides feature.

$$\text{Diagonal: } \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \text{Spherical: } \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \tag{2.4}$$

Table 2.3: Bayesian Model Descriptions

Model	Description
Dots only (light cognitive load)	Attend only to the number of dots
Sides only (light cognitive load)	Attend only to the number of sides
Dots only (heavy cognitive load)	Attend to both features, but infer based only on the number of dots
Sides only (heavy cognitive load)	Attend to both features, but infer based only on the number of sides
Sides and Dots	Inference using both features
Null Model	Guess the stimulus's category randomly on each trial

Bayesian Observer Object Categorization

Our models acknowledge the limitations of sensory perception by adding random noise to Bayesian observers' sensory measurements. Their sensory measurements for dots and sides, S_d and S_s, are given by adding Gaussian noise (ϵ_s, ϵ_d) to true stimulus measurements (ω_s, ω_d) (Equations 2.5 and 2.6). On each trial, a random number was drawn from the ϵ_s and ϵ_d distributions to determine the Bayesian observer's level of sensory imprecision. The variance values which describe the sensory imprecision associated with each model, (ψ_s and ψ_d), were determined experimentally in a different experiment, forthcoming from our lab, which used the same objects (Table 2.4). By adding Gaussian noise which accounts for human participant sensory imprecision to our Bayesian observers, our models are able to more accurately capture the nature of participant performance.

Table 2.4: Bayesian Models Simulated Sensory Imprecision

Model	ψ_s	ψ_d
Dots only (light cognitive load)	N/A	0.78
Sides only (light cognitive load)	1.26	N/A
Dots only (heavy cognitive load)	N/A	1.18
Sides only (heavy cognitive load)	1.65	N/A
Sides and Dots	1.65	1.18
Null Model	N/A	N/A

$$\epsilon_s \sim \mathcal{N}(0, \psi_s^2) \quad \text{and} \quad \epsilon_d \sim \mathcal{N}(0, \psi_d^2) \tag{2.5}$$

$$S_s = \omega_s + \epsilon_s \quad \text{and} \quad S_d = \omega_d + \epsilon_d \tag{2.6}$$

Bayesian observers estimated to which category the current object belongs via maximum likelihood estimate (MLE) based on their current estimates of the parameters which define Elyk and Noek categories. In the case of object categorization, the MLE is equivalent to the *maximum a posteriori* (MAP) estimate because there are equal numbers of Elyk and Noek objects, leading to a uniform prior. For computational convenience, hypotheses were discretized and posteriors were evaluated via numerical solution. Category mean dots and sides (μ_d, μ_s) were each discretized into 5 evenly spaced hypothesized values, and 7 evenly spaced values as hypotheses for each of dots and sides standard deviation (σ_s, σ_d). This gives a total of 1225 mutually exclusive hypotheses for parameters defining Elyk and 1225 mutually exclusive hypotheses for parameters defining Noek. Whichever likelihood was greater in equations (2.7) and (2.8) determined the observer's categorization. If the two likelihoods were equal, the observer would guess at random.

$$P(D|\text{Elyk}) = \sum_{i=1}^{1225} \frac{1}{2\pi\sigma_{d_i}\sigma_{s_i}} \exp\left(-\frac{1}{2}\left(\frac{(S_d-\mu_{d_i})^2}{\sigma_{d_i}^2} + \frac{(S_s-\mu_{s_i})^2}{\sigma_{s_i}^2}\right)\right) \tag{2.7}$$

$$P(D|\text{Noek}) = \sum_{i=1}^{1225} \frac{1}{2\pi\sigma_{d_i}\sigma_{s_i}} \exp\left(-\frac{1}{2}\left(\frac{(S_d-\mu_{d_i})^2}{\sigma_{d_i}^2} + \frac{(S_s-\mu_{s_i})^2}{\sigma_{s_i}^2}\right)\right) \tag{2.8}$$

Learning From Feedback - Parameter Estimation

After receiving feedback on their most recent prediction, Bayesian observers updated their estimates for parameters defining "Elyk" and "Noek." After initial uniformly distributed priors for the first trial, Bayesian updating was used, whereby the prior probability of a parameter vector on trial i is given by the posterior probability of the same parameter vector before trial i, $P(H|D_{<i})$. $D_{<i}$ represents all data collected in all trials before the current one. Equation 2.9 is the probability density of a two-dimensional Gaussian distribution with no covariance, and is the likelihood function of a Bayes-optimal observer for each hypothesis H and sensory data D. H and D are vectors: $H = (\mu_d, \mu_s, \sigma_d, \sigma_d)$, $D = (S_d, S_s)$. For Bayesian observers that used a categorization strategy which considered only sides or only dots, Equation 2.10, the density of a univariate Gaussian distribution, was used as the likelihood function, where m represents the sensory modality used.

$$P(D_i|H) = \frac{1}{2\pi\sigma_d\sigma_s} \exp\left(-\frac{1}{2}\left(\frac{(S_d-\mu_d)^2}{\sigma_d^2} + \frac{(S_s-\mu_s)^2}{\sigma_s^2}\right)\right) \tag{2.9}$$

$$P(D_i|H) = \frac{1}{\sigma_m\sqrt{2\pi}} \exp\left(-\frac{(S_m-\mu_m)^2}{2\sigma_m^2}\right) \tag{2.10}$$

The posterior probability of each hypothesized parameter vector, given the data, was computed using Bayes' Theorem (2.11).

$$P(H|D_{\leq i}) = \frac{P(D_i|H)P(H|D_{<i})}{\sum_{m=1}^{1225} P(D_i|H_m)P(H_m|D_{<i})} \tag{2.11}$$

This posterior probability as determined by (2.11) was calculated for each of the 1225 hypothesized parameter combinations. In this way, the parameters which define the Elyk and Noek distributions could be determined. In order to smooth the random noise added on each trial which accounts for human sensory imprecision, we ran each Bayesian observer 10 times and took the average performance.

Comparing Human Participant to Bayesian Observer

We used Bayesian model comparison to determine which Bayesian observer model aligns most closely with human performance. This quantifies the degree of belief that a participant's performance arose from the strategy a given model performs. If there is a high probability that a participant performed according to a particular model, then it is likely that the participant's actual classification strategy is the same as that model.

For each hypothesized piece, ρ_j, the probability of observing that piece, given that the true category is Elyk, is given by Equation 2.12. The first term of (2.12) represents the probability of drawing piece ρ_j given Elyk's parameters, H; the second term represents the participant's belief that the parameter vector H correctly describes Elyk. The output of Equation 2.11, $P(H|D_{\leq i})$, becomes $P(H|D_{\leq i}, \text{Elyk})$ once the participant is informed that the true category is Elyk. Equations 2.12 through 2.15 describe the computations for Elyk; the process for Noek is the same.

$$P(\rho_j|\text{Elyk}) = \sum_{k=1}^{1225} P(\rho_j|\text{Elyk}, H_k)P(H_k|D_{\leq i}, \text{Elyk}) \tag{2.12}$$

The result of (2.12) is used in Equation 2.13 to find the probability of observing the participant's hypothesized measurement m_k, given that the true category is Elyk. We apply the

law of total probability while conditioning on the true category being Elyk: any object could be part of the Elyk distribution, and a participant's hypothesized measurement could have arisen from any true object. See Appendix B for further mathematical justification of (2.13).

$$P(m_k|\text{Elyk}) = \sum_{j=1}^{25} P(m_k|\rho_j)P(\rho_j|\text{Elyk}) \tag{2.13}$$

Equation 2.13 defines the likelihood of perceiving a certain sides and dots measurement given the true category. This is used in Equation 2.14, Bayes theorem with uniform prior probabilities, to find the posterior probability that Elyk is the correct category given the perceived measurement.

$$P(\text{Elyk}|m_k) = \frac{P(m_k|\text{Elyk})}{P(m_k|\text{Elyk}) + P(m_k|\text{Noek})} \tag{2.14}$$

The probability that the participant reports that piece ρ_g belongs to the Elyk category, $P(\text{“Elyk”}|\rho_g)$ (Equation 2.15), is given by the the probability of any hypothesized measurement, such that Equation 2.14—the probability that the measurement truly belongs to Elyk—is greater than 0.5. This is the probability of any hypothesized measurement, so we use the OR probability rule for independent events to take the sum of all eligible probabilities.

$$P(\text{“Elyk”}|\rho_g) = \sum_k \{Z_k\}, \quad Z_k = \{P(m_k|\rho_g)|P(\text{Elyk}|m_k) > 0.5\} \tag{2.15}$$

The posterior probability of each model, given by Equation 2.16, has a binomial likelihood function, where the binomial coefficient may be omitted because it will appear in the numerator and denominator of Bayes' theorem. $\mathcal{M}_j$ represents the predicted participant classification strategy (one of the six models from Table 2.3). b represents the participant's response on a given trial, and is set to 1 if the human participant categorized the object as Elyk and 0 if they chose Noek. Recall that $1 - P(\text{“Elyk”}|\rho_g) = P(\text{“Noek”}|\rho_g)$ because of the two mutually exclusive response options for participants. n represents the number of trials the participant has completed.

$$P(\mathcal{M}_j|D) = \frac{\prod_{i=1}^{n} P(\text{“Elyk”}|\rho_g)^b P(\text{“Noek”}|\rho_g)^{1-b}}{\sum_{k=1}^{4}\prod_{i=1}^{n} P(\text{“Elyk”}|\rho_g)^b P(\text{“Noek”}|\rho_g)^{1-b}} \tag{2.16}$$

2.3 Temporal Video Segmentation

We collected 54 videos of each participant performing the task. Each video was approximately one minute in duration, and contained 5 trials. With 24 participants, this leaves approximately

1296 hours of video. A forthcoming experiment from our lab has a similar amount of video data which is in the same format. We are interested in what factors correlate with how long participants hold a stimulus before estimating its category on each trial. In particular, we hypothesize participants will: take less time to classify stimuli which have a high probability of belonging to a given category; take more time in experimental conditions with higher category σ values; take more time in the first few blocks of the experiment. We also aim to investigate the correlation between our model's predicted participant classification strategy and the duration of trials. Investigating these questions requires data indicating the duration of trials in as many videos as possible. Manually inspecting each video to determine the duration of each trial is practically infeasible due to the large amount of data we collected. To overcome this challenge, we developed a computer vision program to determine automatically the participant's duration of contact. We are interested in segmenting video into two alternating categories: during a trial, when the participant is touching the stimulus, and between trials, when the experimenter is switching the stimulus to be presented on the next trial. See Figure 2.4 for a visual depiction of the goal of our task.

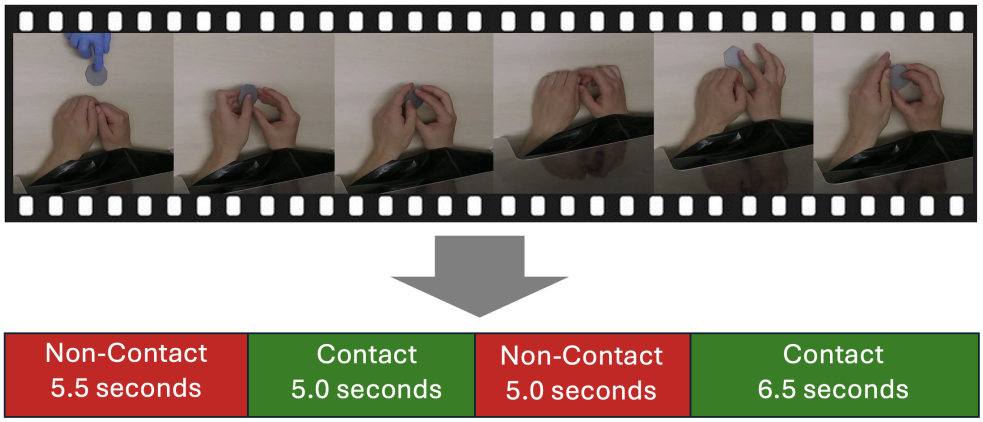


Figure 2.4: The goal of our temporal video segmentation task: the program segments a video of participants performing the haptic categorization experiment in the temporal dimension into alternating actions and identifies the duration of each sub-sequence.

Temporal Video Segmentation Approach

Shan et al., 2020 introduced a dataset with 100,000 annotated images of hands in contact with objects. Their annotations include bounding boxes round the hands and the objects with which they are in contact. They distinguish between stationary objects (e.g. tables and furniture) and portable objects which can be moved by hands. They train a modified version of the popular Faster-RCNN object detection network (Ren et al., 2015) on their dataset, which obtains strong results.

We apply the object detection model developed by Shan et al. (2020) on frames extracted from participant videos. We added a post-processing step to reduce false positives caused by the algorithm detecting the experimenter's hand in contact with the stimulus. If we had not added this step, then the algorithm would falsely report the hand-object contact associated with a participant being mid-trial when the experimenter is swapping the stimulus between trials. We identified frames in which an experimenter was holding the object by leveraging the fact that experimenters wore blue latex gloves. If more than half of the pixels in a bounding box are blue, then we identify that the hand belongs to the experimenter. We determine that a pixel is blue if it has an HSV color code between (90, 50, 50) and (130, 255, 255). See Fig. 2.5 for a visual depiction of the range of "blue" pixels which we associate with experimenter hands in blue latex gloves. Our process for temporally segmenting video into segments with and without participant hand-object contact is described in Algorithm 1.

Algorithm 1 Detect Hand-Object Contact in Video Frames

Input: Video with N frames
  Initialize $contactFrames \leftarrow \emptyset$ ▷ Stores indices of frames with hand-object contact
  for $i = 1 : N$ **do**
    $frame \leftarrow$ extractFrame$(video, i)$
    $detectedItems \leftarrow$ objectDetector$(frame)$ ▷ Object detector created by Shan et al., 2020
    $nonBlueHandDetected \leftarrow$ FALSE
    for all *item* in $detectedItems$ **do** ▷ Exclude blue-gloved hands of experimenter
      if $item.type =$ "Hand" **then**
        $bluePixelPercentage \leftarrow$ bluePixelChecker$(item.boundingBox, frame)$
        if $bluePixelPercentage < 50$ **then**
          $nonBlueHandDetected \leftarrow$ TRUE
          break
        end if
      end if
    end for
    if $nonBlueHandDetected$ **then**
      $contactFrames \leftarrow contactFrames \cup \{i\}$
    end if
  end for
Output: $contactFrames$

Figure 2.5: Gradient of blue colours from HSV color code (90, 50, 50) to (130, 255, 255). This is the range of blue values which we associate with the experimenter's hands in blue latex gloves.

Data Preparation

Initially, the videos were processed to focus on the hands. This was achieved by zooming the original 1920x1080 pixel videos to a 480x480 pixel frame, centered on the participant's hands. We used Google's MediaPipe Python library (Lugaresi et al., 2019) for hand detection. To maintain consistency and avoid jittering effects due to intermittent hand detection, the center of the first frame where both hands were visible was used as a reference frame. All frames were zoomed to this reference, so that the center of the hands in the reference frame is the center of all processed frames.

After we ran multiple participants through the procedure, we encountered technical difficulties with our video recording setup. We switched the angle of the recording slightly (from top-down to angled top-down), and switched our recording to 60 frames per second (fps) from 30 fps. The difference between camera setups is subtle, but visually noticeable; Figures 2.3a and 2.3b show the subtle differences. The original data collection was conducted using a GoPro Hero Session camera, capturing footage from 11 participants across both days and 6 participants exclusively on the first day. Subsequent recordings were made using a GoPro Hero Session 5 camera.

Data Labelling

While our method does not require domain-specific labeled data, we did need to label a portion of our data so that we could compare our model's predictions to a known ground truth. We focused only on assigning a "contact" or "non-contact" label to each frame, rather than drawing bounding boxes around objects within each frame. While the latter method would allow a higher degree of precision in our evaluation, the former is much less time-consuming and enables us to compare our model's final predicted label on each frame to its true label.

The data labelling process involved manually identifying transition frames: those depicting the initiation and cessation of contact between the hand and the object. This identification was carried out through manual visual inspection. Frames between these transitions were automatically sorted into their respective categories.

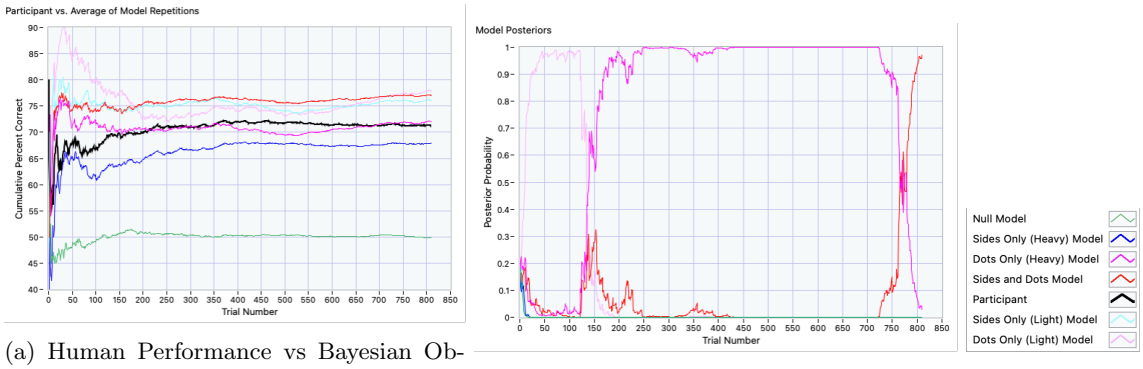
Our transition labeling process resulted in 87,352 labeled frames. These frames came from videos of three volunteers performing a single block of the task. Each video was approximately 10 minutes in duration. The volunteers were not part of the main study. One video was filmed with the initial camera setup, and the other two videos were filmed with the modified camera angle.

During the recording of both videos with the modified camera setup, the camera angle was slightly adjusted a few times. This was done to increase the diversity of the training data to better reflect small differences in camera angle which might occur in the experimental data.

Results

3.1 Experimental and Modelling Results

Human performance is similar to Bayesian observer models. Figure 3.1a shows performance of a human participant on both days of the experiment compared with all Bayesian observers. Figure 3.1b shows the participant's Bayesian model comparison results over the course of the same trials to estimate categorization strategy. For this participant, the model which best describes the participant's strategy fluctuates over the course of the experiment. In approximately the first 3 blocks, or 135 trials, our model indicates that the participant uses a dots only strategy with a light cognitive load: this posits that they do not attend to the number of sides. For most of the remaining experiment, we estimate that they switch to the dots only strategy with a high cognitive load: they attend to both sides and dots, but only make an inference using the number of dots. The performance in trials in which we estimate they use the dots only with a heavy cognitive load (Figure 3.1b) is similar to that of a Bayesian observer which simulates this strategy (Figure 3.1a). Around the last block of Day 2, our model estimates that they switch to optimally combining information about the number of sides and dots.



(a) Human Performance vs Bayesian Observers

(b) Model Comparison

(c) Legend

Figure 3.1: (A) Sample participant asymptotic performance is similar to Bayesian observer asymptotic performance. (B) Model comparison determines which strategy the human participant is using on a trial-by-trial basis. The most probable strategy switches from incorporating sides only to incorporating both sides and dots.

Our modelling indicates that at the end of each experimental day, most participants used one of two models: the dots only model with a light cognitive load (where they do not attend to the sides feature at all) or the sides and dots model (where they optimally combine information from both stimulus features). More participants used these two strategies after Day 2 (Figure 3.2). This

was determined by reporting the model with the highest posterior probability from our model comparison on the last trial of each experimental day.

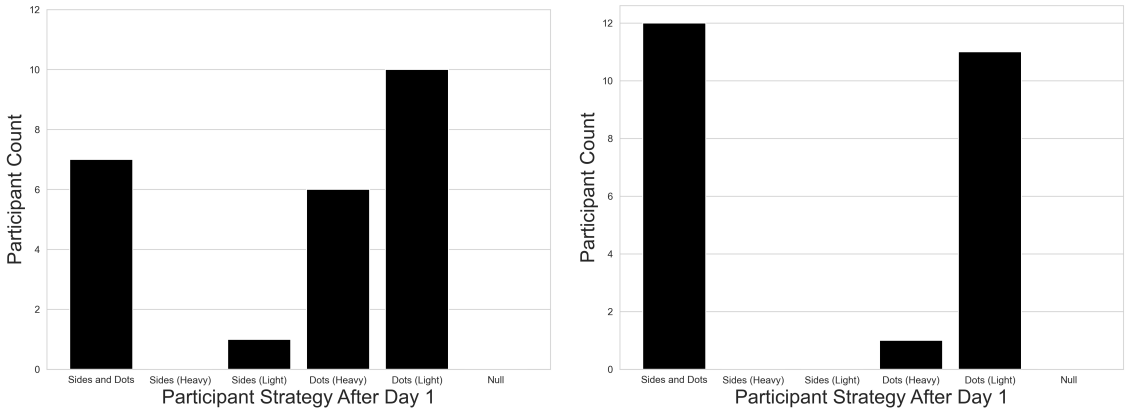


Figure 3.2: Number of participants who ended the experiment with a given strategy, according to our model's prediction after (A) Day 1 and (B) Day 2.

Figure 3.3 illustrates the specific changes in participant strategy from the end of Day 1 to the end of Day 2 (the differences between Figure 3.2a and 3.2b).

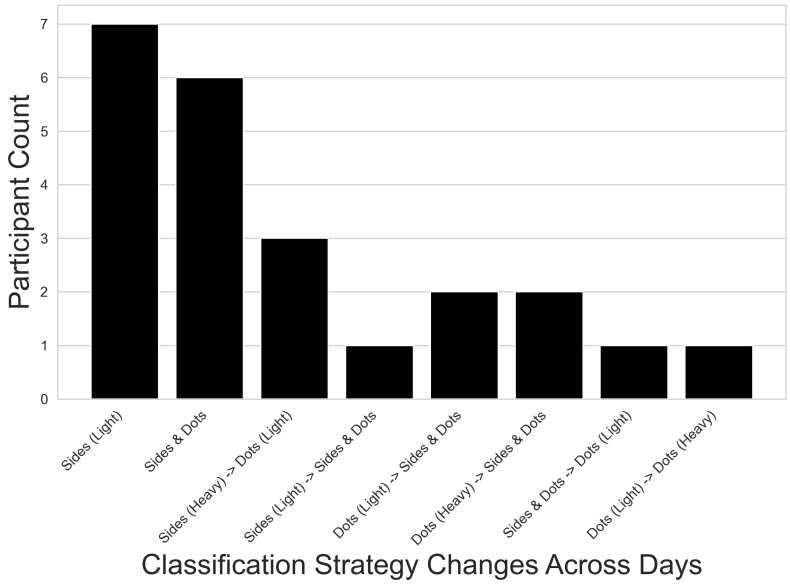


Figure 3.3: Participant Classification Strategy Changes from the end of Day 1 to the end of Day 2. Many participants switched to either the sides only (light) strategy or the sides & dots strategy.

In all four experimental conditions, participants have higher classification accuracy on Day 2 compared to Day 1 (Fig. 3.4). The performance improvement across days is more pronounced for conditions with lower σ_E and σ_N. Groups with $\sigma_E = 0.75$, $\sigma_N = 1.25$ and conditions with $\sigma_E = 1.25$, $\sigma_N = 0.75$ saw similar performance and similar improvement in performance across days.

Notably, the difference between those two conditions is less pronounced than any other pairwise differences between conditions. This is expected because the only difference between these experimental conditions is which label (Elyk or Noek) we assign to the different category σ conditions. The conditions with lower category σ values had higher classification accuracy on both experimental days.

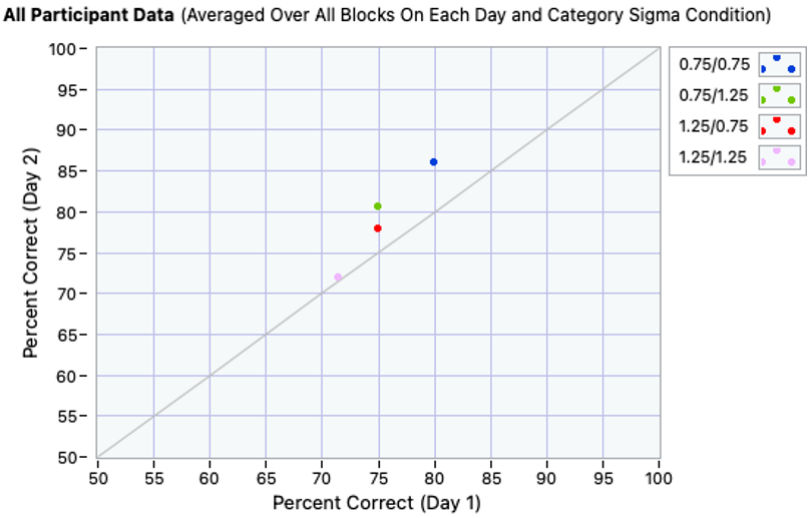


Figure 3.4: Performance improvement across days. "0.75/1.25" denotes $\sigma_E = 0.75$, $\sigma_N = 1.25$

3.2 Temporal Video Segmentation Results

We provide both a qualitative and quantitative analysis of our temporal video segmentation results.

Qualitative Results

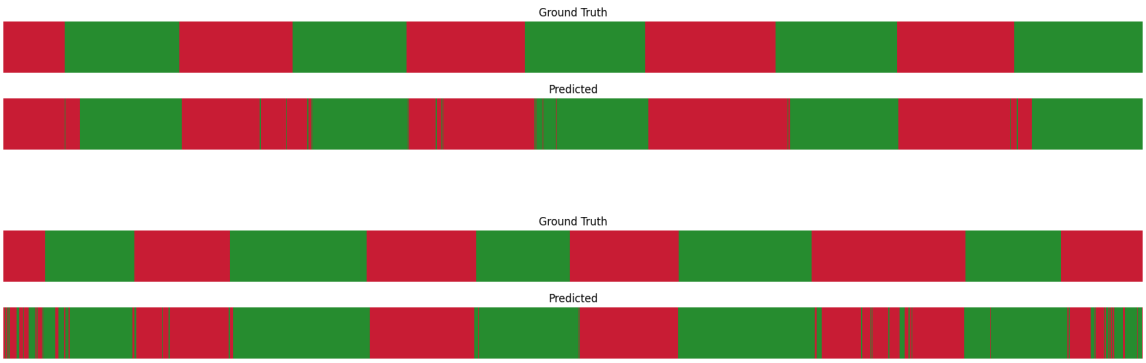


Figure 3.5: Model Predictions Compared to Ground Truth on Two Sample Participants

In Figure 3.5, the green and red segments represent timesteps at which the model predicts there is participant hand-object contact, and where there is none, respectively (code to generate Figure 3.5 was adopted from Lea et al., 2017). We compare the predictions to the ground truth, which we obtained by manually labelling the video frames. Upon visual inspection, it is clear that there is strong qualitative performance.

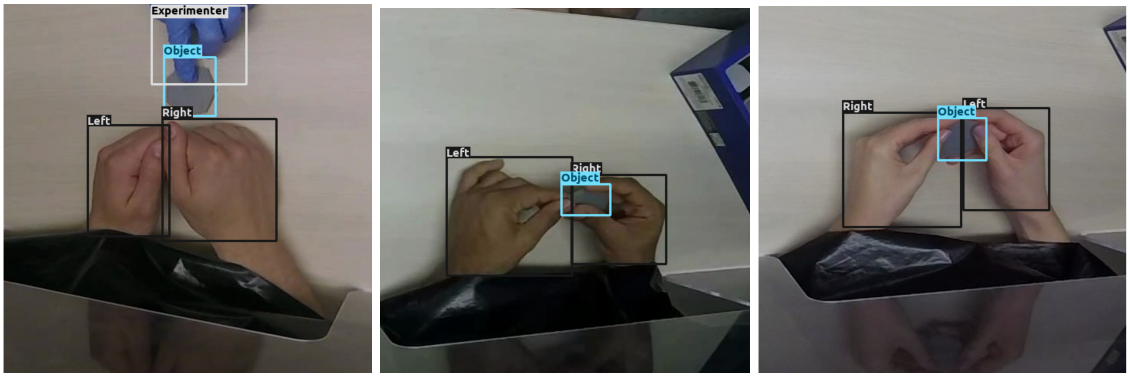


Figure 3.6: Sample frames in which the object detection model performs well

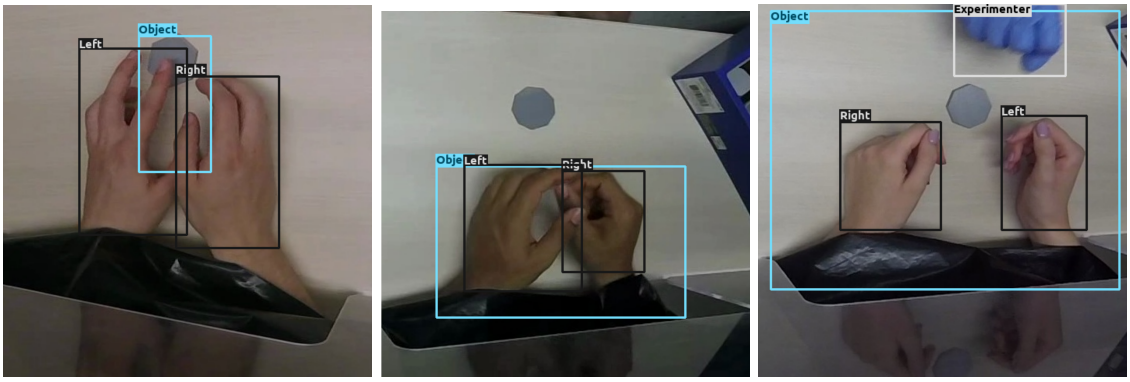


Figure 3.7: Sample frames in which the object detection model performs poorly

Figures 3.6 and 3.7 highlight examples of the object detection model performing as intended and making mistakes, respectivley. Additional qualitative results can be found in Appendix C.

Quantitative Results

There are three standard metrics used in the temporal action segmentation literature (Ding et al., 2024): Mean over Frames (MoF), Segmental F1 score (F1@τ) (Lea et al., 2017), and Edit Score. We evaluated our model using these metrics on the videos of three volunteers, who each completed one block of the experiment which consisted of 87,352 frames. Each of the metrics range from 0 to 100, with 100 being the best score.

The MoF metric, defined as:

$$\text{MoF} = 100 \cdot \frac{\text{\# of correctly labeled frames}}{\text{\# of incorrectly labeled frames}} \tag{3.1}$$

serves as a straightforward measure of frame-by-frame accuracy, disregarding the temporal structure of action segments.

In contrast, the Segmental F1 score (F1@τ) is an evaluation metric which accounts for temporal consistency. This metric extends the traditional F1 score (which is the harmonic mean of precision and recall) by considering the temporal overlap between predicted and ground truth segments. For a predicted segment to be considered a true positive (TP), it must intersect with a ground truth segment with an Intersection over Union (IoU) of at least τ. Therefore, this metric evaluates both the presence and the precise temporal localization of actions:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.2}$$

$$\text{F1} = 100 \cdot \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Precision}} \tag{3.3}$$

The Edit Score complements these metrics by evaluating the sequence of predicted actions against the ground truth, considering the order of actions and penalizing for insertions, deletions, and substitutions necessary to align the two sequences. It is calculated using normalized Levenshtein distance between the two action sequences, with a higher score indicating fewer required modifications and, consequently, a more accurate model:

$$\text{Edit Score} = 100 \cdot \left(1 - \frac{\text{Levenshtein distance(Predicted Labels, True Labels)}}{\text{Length of Sequence}}\right) \tag{3.4}$$

The Levenshtein distance component of the edit score equation measures the minimum number of edits needed to transform one sequence into another, with edits being insertions, deletions, or substitutions of single elements. This distance measure is used to quantify the degree of similarity in two sequences.

The overall MoF across all participants was 87%. This means that 87% of the 87,352 frames were assigned the correct label. Across all tested videos, we found strong quantitative results in the MoF metric, and poor results in the temporally-aware F1@τ and edit score metrics. This discrepancy in performance indicates that the present model is not suitable for unsupervised

Table 3.1: Temporal Video Segmentation Quantitative Results

Metric	Volunteer 1	Volunteer 2	Volunteer 3
MoF	69	89	92
Edit Score	18	7	12
F1@10	24	13	22
F1@25	23	11	22
F1@50	16	9	19

deployment on 1000+ hours of video; however, the high MoF score and strong qualitative results indicate that the present model is a promising system which delivers strong preliminary results. In the Discussion section, we discuss in depth how we expect that we could leverage the high MoF scores to improve the other quantitative evaluation metrics.

Discussion

Bayesian Modelling of the Haptic Categorization Task

We compared human performance on a haptic categorization task to that of a series of Bayesian observers so that we could understand what strategy participants used when performing the task. We conceptualize the behavior of participants engaged in our haptic categorization task as employing a Bayesian approach to categorization. This approach involves learning the parameters that delineate two newly introduced probabilistic categories, Elyk and Noek, which are characterized by 2-dimensional Gaussian distributions with spherical covariance structure. In different experimental conditions, we alter the variance of the categories to investigate how this impacts participant strategy. Experimental conditions in which there is a wider variance are inherently more challenging: any classification strategy which must learn the parameters of two distributions which heavily overlap has a lower theoretical performance compared to when there is minimal overlap between the distributions.

Figure 3.2 illustrates that participants most often use the strategy of attending only to the number of dots, or the strategy which entails optimally combining information about the number of both sides and dots. This finding remained true across all experimental conditions. These are the two strategies which yield the highest performance in Bayesian observers. While the strategy of attending only to dots sacrifices information about the number of sides, it is still a strong strategy because it minimizes the sensory imprecision on each trial (Table 2.4). In this way, attending only to dots makes up for the lack of information about the number of sides by maximizing the precision of this limited information. As a result, the strategy of attending only to dots can sometimes result in more accurate classifications than attending to both sides and dots.

The fact that attending to only one of two tactile features can be optimal is at first counter-intuitive: there are some stimuli for which the sides feature is informative and the dots feature is not. For these stimuli, the number of dots indicates that the stimulus has an equal probability of belonging to each category, but the number of sides does not (Figure 4.1). On trials with these stimuli, attending only to the number of dots would lead a participant to guess the stimulus' category at random; attending to both features would allow the participant to produce a well-informed inference using the informative sides feature. Attending to both features gives participants more information, facilitating use of whatever information is most useful on a given trial. The reason that obtaining this additional information is not necessarily optimal is that the

process acquiring it comes at the cost of the precision of all information.

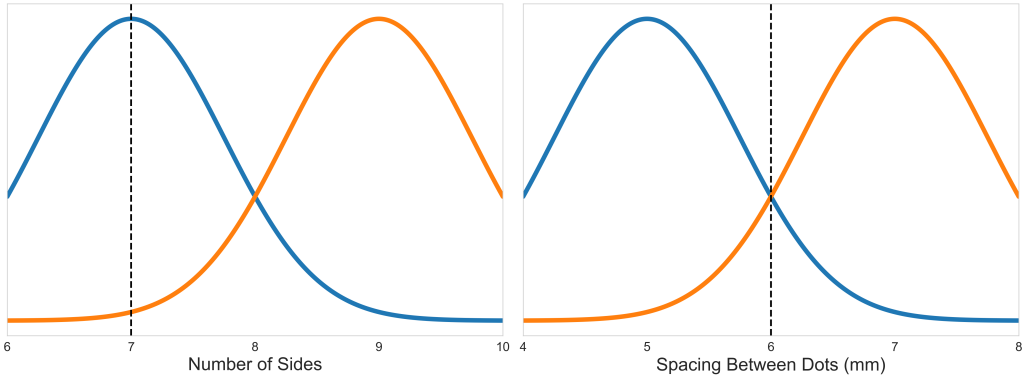


Figure 4.1: Example of conditions under which combining both cues to form inference is helpful. The blue and orange distributions represent the Elyk and Noek categories, respectively. The black dotted line represents the features of a stimulus which is informative in its number of sides but not its dot spacing.

Unlike the above example, attending to only one feature provides sufficient information about the stimulus to form an accurate inference about its category in most trials. This is necessarily true because stimuli near a category's mean are presented to participants and Bayesian observers more frequently. In these trials, the lower sensory precision associated with attending to an additional feature can diminish performance enough to result in an incorrect classification; on the same trial, the extra precision afforded by attending only to the number of dots could have resulted in a correct classification. Throughout the rest of this discussion, we refer to the strategy of 1) attending only to dots and 2) attending to sides and dots as the "dominant strategies."

Many participants who were not using one of the two dominant strategies after Day 1 ended Day 2 using one of them (Figures 3.2 and 3.3). This switching of strategy is associated with an improvement in performance across days (Figure 3.4). Participants switching strategies over the course of the experiment is expected: if they try a new strategy for a few trials, they receive rapid feedback which indicates whether they should continue with the new strategy or revert to the previous one. In particular, this rapid feedback is more likely to affect participant strategy in experimental conditions in which the Elyk and Noek categories have narrower distributions (smaller category σ). Under these conditions, participants are more likely to receive positive feedback when switching to a more optimal strategy. In the more challenging experimental conditions where the category distributions are wider, there is a comparatively high probability that on a given trial, a participant's optimal inference will be incorrect: a stimulus which is more likely to belong to the Elyk could be identified as a Noek on multiple trials due to chance. If this

were to happen when a participant is in the process of evaluating a new strategy, it could lead them to think the strategy is at fault, rather than the challenging nature of their experimental condition. The probability of a participant encountering such a challenging situation is lower when the category distributions are narrower. It is therefore more challenging for participants to evaluate a given strategy in conditions with wider category variance.

In practice, human participants are likely to account for cognitive load when picking their classification strategy. In other words, we expect participants to pick strategies which both maximize task accuracy and minimize cognitive load. In Section 2, we discuss the high cognitive load associated with attending to both tactile features. Beyond the high cognitive load associated with attending to multiple features, we hypothesize that there is a higher cognitive load associated with feeling the number of sides of our stimuli compared to the number of dots. This is reflected in the higher sensory imprecision for this strategy which was quantified in a forthcoming experiment from our lab (Table 2.4). This hypothesis is also consistent in the context of the present results (Figure 3.3): almost all participants who ended Day 1 with a strategy which demands a higher cognitive load switched strategies over the course of Day 2. The significance of Figure 3.4 can also be highlighted in the context of minimizing cognitive load and maximizing accuracy: we expect that strategies which rely on the number of sides for inferring object category are used less frequently because these strategies yield lower accuracy and demand a higher cognitive load. After Day 1, our model estimated that there were few participants who relied exclusively on the number of sides in forming their inferences; after Day 2, most of these participants switched to one of the dominant strategies (Figures 3.2 and 3.3).

Despite these promising results, we are aware of some limitations of our experiment. Our stimulus set includes only 25 objects. This has a few undesirable effects: two very similar objects are relatively far from one another within the category's probability distribution, and participants might be able to memorize characteristics of specific objects over the course of the experiment. The latter point is most pronounced in conditions with small category σ values: in most trials under these conditions, participants are presented with one of 9 most probable Elyk or one of 9 most probable Noek objects. This could facilitate stimulus-category memorization, rather than learning via Bayesian parameter estimation.

Participant attention is another possible limitation of our study. Participants are likely to forget some information about the categories between days 1 and 2, but our model makes no distinction between days. Our participants often had weaker performance on the first block of Day 2 compared to the last block of Day 1; our Bayesian observers had no such limitation.

Additionally, each experimental day took close to 2 hours to complete. It is certainly possible that participant attention span and effort decreased over the course of our study.

Going forward, we would like to extend our experiment in two ways. First, we would like to investigate the effect of changing the proportion of Elyk and Noek stimuli on performance. In the present study, Elyk and Noek objects are presented to the participant in equal portions. Changing this proportion without informing participants of its new value could facilitate further analysis of the Bayesian models we present here. Changing this proportion would have the effect of changing the prior probability for a category when participants and observers classify objects on each trial.

We would also like to reproduce the experiment with a different set of stimuli. The number of sides feature in the current stimulus set is a discrete feature which we describe using a continuous probability density. A new stimulus set which replaces this feature with a continuous one, such as spherical object size, for example, could map more naturally onto the continuous Gaussian densities which define our categories. Ideally, a future stimulus set would include more total stimuli, so as to avoid the memorization limitation discussed above.

Temporal Video Segmentation

Our temporal video segmentation algorithm yields strong preliminary results. The results are considered strong as preliminary results primarily because of the high MoF score (Table 3.1). While the temporally-sensitive F1@τ and edit score metrics are low, this is largely the result of frequent over-segmentation errors (Figure 3.5), which are discussed in depth below. We expect that we can leverage the high MoF score to improve our system in a way which alleviates the over-segmentation errors that cause poor performance in temporally-sensitive quantitative metrics.

The approach described above is our second attempt at designing an algorithm for our purposes. Appendix A discusses our initial approach to video segmentation which yielded unsatisfactory results. The backbone of our final successful algorithm is an object detection model trained on a diverse dataset with annotated hands in contact with objects from across the internet (Shan et al., 2020). Use of the publicly available dataset and object detection model enabled us to achieve our strong results without any manual data labeling.

We identified two types of common errors which our model makes:

- Common Error 1: Overestimation of the duration between trials, where the model does not detect that the participant is touching the object until they have lifted it off of the table.
- Common Error 2: Over-segmentation errors, where there are very short time periods–as

short as one or two frames–where the model's predicted contact state switches.

We discuss potential causes of these two types of error, and propose a series of extensions to the current work which could ameliorate these errors, below.

First, we discuss Common Error 1. By qualitatively examining predicted bounding boxes in mislabeled frames, we determined that this phenomenon occurs because the model detects the participants' hands resting on the table, but does not properly detect the stimulus which is on the table; rather, it predicts that the stimulus is part of the table while the participant has not lifted the stimulus off of it. The object detection model distinguishes between portable objects (like our stimuli) and stationary objects (like the table on which the participants rest their hands) (Shan et al., 2020). When the stimulus is still on the table and the participant is touching it, the object detection model can often draw a bounding box around the table, and predict that the participant is touching a stationary object: it fails to identify that the participant is touching a thin portable object which is itself in contact with a stationary object. To ameliorate Common Error 1, we could train an object-detection algorithm specifically designed to recognize our stimuli. In this way, we could detect when there is overlap between a detected hand and a detected stimulus, facilitating the inference that there is hand-object contact in that frame. Algorithm 2 describes how this approach could detect hand-object contact. We expect that this method would have the highest precision and accuracy when the stimulus is still on the table because this is when the stimulus has the least visual obstruction.

Algorithm 2 Contact Detection using Only Hand and Stimulus Detector

Input: Video with T frames, handDetector, stimulusDetector
  Initialize $contactFrames \leftarrow \emptyset$ ▷ Stores frames where hand contacts stimulus
  for $t = 1 : T$ **do**
    $frame \leftarrow$ extractFrame$(Video, t)$
    $handBoxes \leftarrow$ handDetector$(frame)$
    $stimulusBoxes \leftarrow$ stimulusDetector$(frame)$
    for all $hBox \in handBoxes$ **do**
      for all $sBox \in stimulusBoxes$ **do**
        if isOverlap$(hBox, sBox)$ **then**
          $contactFrames \leftarrow contactFrames \cup \{t\}$
        end if
      end for
    end for
  end for
Output: Frames where hand contacts stimulus: $contactFrames$

Creating training data for an object detection model is usually time-consuming; however, our current model can already accurately identify our stimuli in many situations. The stimuli which

are detected by the current object detection algorithm can be used as training data for a stimulus-specific object detector (Figure 3.6). We can use heuristics to improve the quality of these automatically generated labels to ensure that only quality samples are used to train this domain-specific object detector. We could only accept bounding boxes for stimuli which have similar length and width (this would sidestep the error in Figure 3.7a); we could also require that the bounding box around a detected object be smaller than hands detected in the frame. These two rules would prevent the erroneous bounding boxes in Figure 3.7 from being identified as training data for this object detector.

We now discuss Common Error 2: over-segmentation errors. Over-segmentation errors are well-established as a prominent challenge in temporal action segmentation literature (Ishikawa et al., 2021; Xu et al., 2022; Ding et al., 2024). In our case, some over-segmentation errors arise because our algorithm is not temporally-aware: it classifies each frame individually, without taking account of context from surrounding frames. An ideal system would be able to recognize that if a participant picks up the object, and manipulates it in a way which causes it to be briefly hidden from the camera, then they probably did not drop the object and pick it back up very quickly. Temporally-aware computer vision models do exist for tasks similar to ours; we opted not to implement them because most of these models are fully-supervised, meaning they require a large amount of domain-specific labeled data (Ding et al., 2024). We would not have been able to leverage the broad dataset from Shan et al. (2020) if we were to use one of these models.

In the temporal action segmentation literature, "timestamp supervision" refers to a specific type of semi-supervised learning, where most training data is unlabeled, but in each video, there is (at least) one frame with a class label and the label's associated timestamp (Li et al., 2021, Ding et al., 2024). Obtaining training data for systems trained using timestamp supervision is significantly less time-consuming than for a fully-supervised system. While we did consider using a model which required only timestamp supervision, this still would have required labeling additional training data. Part of our goal in this project was to minimize the amount of time spent manually labeling data, which is why we chose not to use one of these systems.

We propose a technique which could be implemented in the future which leverages our current model in conjunction with a temporally aware machine learning model in order to minimize over-segmentation errors. We propose that a temporal action segmentation model which requires timestamp supervision could be trained on a dataset that is automatically generated using our current system. A dataset with timestamp supervision does not require labels on every frame in videos which serve as training data. This is a key insight, and allows us to extract a small number

of labels from our system's predictions to form this dataset. We would have our current model predict the class of every frame in the entire dataset, and extract only the frames where we are most confident about the model's predicted labels. This would be done on video for which we have no ground truth labels. Figure 4.2 shows both our model's predicted labels and ground truth labels from a part of a video. The black dotted lines represent frames about which we have the highest prediction confidence because they are in the middle of a sequence of a single predicted class. The frames which correspond to the black dotted lines could be used in a dataset for timestamp supervision. In this way, we could automatically generate a dataset for a temporally-aware machine learning model without manually labeling any data. Over-segmentation errors from our current model are unlikely to meaningfully affect this method of generating training data. This is important because these over-segmentation errors which this technique sidesteps are a primary factor which influence our model's poor performance in temporally-aware evaluation metrics.

Figure 4.2: Black dotted lines represent frames for which we are very confident that their label is accurate because they are surrounded by many frames with the same prediction.

This automatically-generated dataset could be used to train a temporal action segmentation model designed to learn from timestamp supervision such as that from Li et al. (2021). We expect that by training a temporally-aware video segmentation model in this way, we could significantly reduce the number of over-segmentation errors caused. This reduction in over-segmentation errors should result in an improvement in F1@τ and edit score metrics. We also expect that we could use the predictions from this temporally-aware model to quantify which factors impact how long participants take to classify our objects during the haptic categorization task.

Conclusion

The fact that most participants used one of the dominant strategies across all experimental conditions indicates that learning using only the tactile system is Bayes-optimal under a variety of circumstances. This remains true even when the haptic categorization task is made more challenging. We designed a computer vision system which can segment video into temporally distinct segments based on when participants are holding stimuli. The system did not require manually labeled training data in order to obtain promising preliminary results. While the system has high qualitative accuracy and high frame-wise accuracy, it is prone to over-segmentation errors, which diminish its quantitative performance on temporally-aware evaluation metrics. We propose a method which we expect would address these over-segmentation errors in future work without the need for manual data labeling. We expect that engineering a system in this way would allow us to answer precise questions about what factors correlate with how long participants take to classify objects on each trial.

Appendix A: Initial Attempt at Temporal Video Segmentation

Before our current video segmentation approach, we tried a different approach which yielded unsatisfactory results. The approach involved using an image classifier to generate frame-wise probabilities of hand-object contact in each frame and segmenting the time series of probabilities.

Overview of Approach

Algorithm 3 describes an overview of this initial approach. Rather than using object detection to identify key components of an image to make a downstream inference about the entire image like Algorithm 1, this method relies on an image classifier: entire images are classified in a single step. The idea here was that the importance of hand-object contact could be learned by the image classifier with enough labeled training data. The image classifier is trained on manually labeled domain-specific video from participants.

Algorithm 3 Initial Algorithm

Input: Video with T frames
  Initialize $y \leftarrow \emptyset$ ▷ Stores frame-wise contact probabilities
  for $t = 1 : T$ **do**
    $frame \leftarrow \text{extractFrame}(video, t)$
    $p_t \leftarrow$ assigned from image classifier ▷ Current frame contact probability
    $y[t] \leftarrow p_t$
  end for
  Apply changepoint detection on $\{y_t\}_{t=1}^{T}$ time series
Output: set of estimated changepoint indexes

Primer on Image Classification

Deep Learning, or machine learning using deep Artificial Neural Networks, is very popular in modern machine learning, particularly in tasks involving image or text data (LeCun et al., 2015). Supervised Deep Learning models are trained to make predictions based on labeled input data; the training process involves iteratively making predictions while adjusting the network's internal parameters to minimize prediction errors. Convolutional neural networks (CNNs) have been used extensively in image classification tasks, where they achieve state-of-the art results (Chen et al.,

2022; LeCun et al., 2015; Krizhevsky et al., 2012). Deep transfer learning is a technique in deep learning where a neural network developed for a specific task is repurposed as the starting point for a model on a second task (Farahani et al., 2021; Kim et al., 2022). This approach is particularly beneficial in scenarios where the desired task has limited data available for training, and has been used extensively for image classification with CNNs (Ribani and Marengoni, 2019). By leveraging a model which performs well on a diverse dataset, transfer learning allows for the utilization of knowledge (in the form of neural network parameters) which was acquired from the original task; this enhances learning efficiency and performance on the new task. Pre-trained CNNs for transfer learning often leverage parameters from the architecture trained to classify photos from ImageNet, a collection of over 1,200,000 images belonging to 1000 classes (Deng et al., 2009; Ribani and Marengoni, 2019). Like most modern TAS techniques, we use a Deep Learning image classifier as the backbone of our TAS model (Ding et al., 2024).

Image Classification Neural Network

We used the EfficientNetV2-L CNN architecture pre-trained on ImageNet as our base model for transfer learning (Tan and Le, 2021). The training images were labeled into either "hand-object contact" or "non-contact" categories. A neural network's loss function quantifies the quality of a network's predictions: the training process aims to find the parameters which minimize the loss function, thereby maximizing the quality of predictions on unlabeled data. The network was trained to identify the class of these images using Binary Cross Entropy Loss (Equation A.1), where y represents the true binary label of the image and $\hat{y}$ represents the predicted probability of the image belonging to the positive class. The image classification neural network was trained on a single Nvidia GeForce RTX 3090 GPU.

$$\mathcal{L}(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right) \tag{A.1}$$

Time Series Segmentation

The time series representing the probability that each frame contains hand-object contact can be segmented into different groups to determine the duration of each trial. This problem can be conceptualized as one of time series changepoint detection. Specifically, offline changepoint detection techniques can be leveraged, where "offline" refers to the fact that there are no real-time updates to the data we are segmenting. Formally, we want to segment a univariate time series

$y = \{y_1, y_2, \ldots, y_T\}$ with K^* unknown changepoints $\{\tau_1 < \tau_2 < \ldots < \tau_{K^*}\}$, where $y_i \in \mathbb{R}$, $\tau_i \in \mathbb{N}$ (Truong et al., 2020). Each component y_i represents the probability that a given frame has hand-object contact, as determined by the deep image classifier. The binary nature of our segmentation task entails that the structure of the time series is described by Equation A.2, where every segment between changepoints comes from one of two alternating stationary distributions. The first distribution represents the participant touching the object, and the second distribution represents them not touching it. If the image classifier is effective, then we can expect that the mean of Stationary Distribution 1 is near 1 and the mean of Stationary Distribution 2 is near 0.

$$y(t) \sim \begin{cases} \text{Stationary Distribution 1} & \text{if } 1 \leq t \leq \tau_1 \\ \text{Stationary Distribution 2} & \text{if } \tau_1 < t \leq \tau_2 \\ \text{Stationary Distribution 1} & \text{if } \tau_2 < t \leq \tau_3 \\ \text{Stationary Distribution 2} & \text{if } \tau_3 < t \leq \tau_4 \\ \vdots & \vdots \end{cases} \tag{A.2}$$

While the number of changepoints in each video should be 10 (each video contains 5 trials, and each trial contains one transition to begin hand-object contact and one transition to finish it), there are some videos with only 9 changepoints due to delays in the starting of video recording. Thus, we do not know the exact number of changepoints K^*, but we do know that $K^* \in \{9, 10\}$.

The Pruned Exact Linear Time (PELT) algorithm is a technique for detecting the location of changepoints when the number of changepoints K^* is unknown (Algorithm 4, pseudocode reproduced from Truong et al., 2020). Given a user-specified cost function and penalty constant β, PELT is guaranteed to converge to the global minimum cost on the order of $\mathcal{O}(T)$ computational complexity (Killick et al., 2012).

Algorithm 4 PELT

Input: signal $\{y_t\}_{t=1}^T$ with $y_t \in [0, 1]$, cost function $c(\cdot)$, penalty value β
  Initialize Z a $(T+1)$-long array; $Z[0] \leftarrow -\beta$
  Initialize $L[0] \leftarrow \emptyset$
  Initialize $\chi \leftarrow \{0\}$ ▷ Admissible indexes
  for $t = 1 : T$ **do**
    $\hat{t} \leftarrow \arg\min_{s \in \chi}[Z[s] + c(y_{s..t}) + \beta]$
    $Z[t] \leftarrow [Z[\hat{t}] + c(y_{\hat{t}..t}) + \beta]$
    $L[t] \leftarrow L[\hat{t}] \cup \{\hat{t}\}$
    $\chi \leftarrow \{s \in \chi : Z[s] + c(y_{s..t}) \leq Z[t]\} \cup \{t\}$
  end for
Output: set $L[T]$ of estimated changepoint indexes

We implemented the PELT algorithm provided in the Ruptures Python package (Truong et al., 2020). We chose to use this algorithm because it has much lower computational complexity than any techniques for known K^*, making it much faster for prototyping, especially on long videos. In future refinements which rely on time series changepoint detection, we may decide to use Dynamic Programming as our segmentation algorithm, which requires known K^*, and no β parameter.

Results: Image Classification

Image Classifier 1: Trained on Subset of Labeled Dataset

We trained our image classifier on two sets of data. The first was a subset of our labeled dataset which we used for initial proof-of-concept testing. This data was generated using training data consisting of 6068 frames from a single video of one participant using the first camera setup. Figure 2.3a is a sample image in this first dataset, but Figure 2.3b is not. Training on this small amount of data was done only for proof-of-concept purposes. The network was tested on 1516 randomly drawn samples from the same video dataset which were not used in training. This network achieved accuracy and F1 score of 96% (Equation 3.3).

This classifier performed well on extended sections of video which were not in the training set, but did not generalize well to videos of other participants (Figure A.1). This poor generalization is unsurprising because the only training data used was from a single participant and a single camera angle: this biased training data prevents the model from being robust to differences across participants such as classification strategy, skin colour or camera angle.

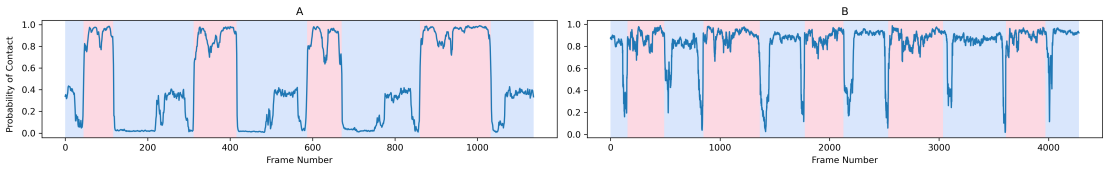


Figure A.1: Shaded regions represent ground truth labels: blue regions represent segments with no contact, red regions represent regions with contact. (A) Probability of contact in video taken from the same source as the training data for the image classifier 1. The classifier was not trained on this portion of video. (B) Image classifier 1 performing inference on video from a participant who is not in the training data.

Image Classifier 2: Trained on Full Labeled Dataset

We aimed to ameliorate this concern about generalization performance with the second dataset. This dataset includes all labeled data, which consists of 87,352 labeled images. A random sample (without replacement) including 80% of the labeled data was used for training, and the rest

was used to evaluate performance. The image classifier achieved accuracy and F1 score of 95% on this larger, more diverse dataset. Despite strong performance on the diverse dataset, the image classifier generalized very poorly to videos of other participants (Figure A.2). At this point, it is unclear why the model did not generalize well. It is possible that the reason is as simple as not enough variation in the training data.

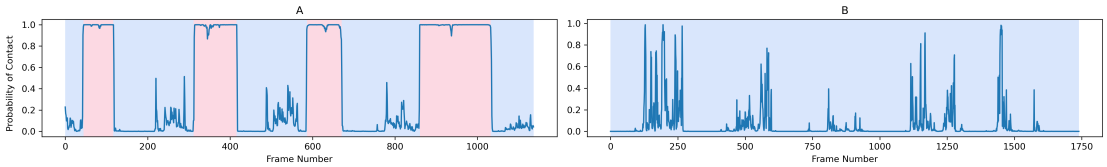
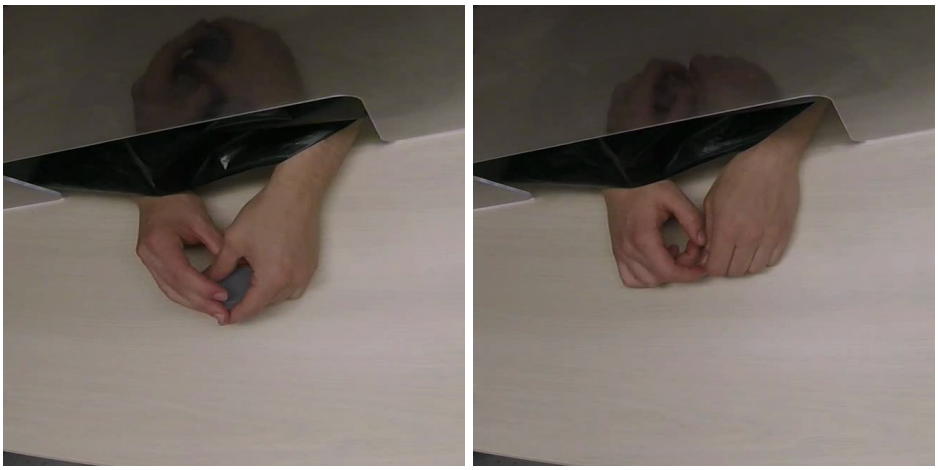


Figure A.2: (A) Probability of contact in video taken from the same source as the training data. The input video is the same as in Figure A.1a. (B) Image classifier 2 performing inference on video from a participant who is not in the training data. Regions in (B) are not shaded according to true labels because this section of data is unlabeled. It is clear that the image classifier does not perform well on this data.

Spikes in probability during non-contact temporal regions coincide with ambiguous frames, such as those in Figure A.3. In some frames when the participant is holding the object, the object is not visible at all from the camera's top-down view. This initial method does not account for temporal context in image classification: each frame is classified independent of the context of previous and future frames. Image classification with temporal context is possible (using I3D feature extraction developed by Carreira and Zisserman (2017)); however, our very poor generalization results even on single images indicates that a method which relies on image classification us unlikely to produce strong results.



(a) Object in Hand (b) Resting Hands Between Trials

Figure A.3: Demonstration of a challenge in image classification for our task: on the scale of the whole image, the differences between (A) and (B) are minimal, but there is an object in the volunteer's hand in (A) but not in (B)

Results: Changepoint Detection

We tested the Pruned Exact Linear Time (PELT) algorithm for changepoint detection using the first image classification model. The dataset we used for testing was the same dataset the image classifier was trained on, but we tested the algorithm close to four minutes of additional video which was not in the training set. In this portion of the video, the participant and camera angle were the same as in the training data. We found that the PELT algorithm using univariate quadratic error loss (Equation A.3) and a penalization value of $\beta = 10$ yielded strong empirical results in this test. In the 8 minute video in which the first half was training data for the image classifier, PELT correctly identified the number of changepoints, and had only one error greater than a quarter-second in a time series with 92 changepoints. This is a very promising result, and we therefore expect that if an image classifier performs well on new data, then PELT will be a suitable algorithm to segment the time series of probabilities which the image classifier outputs.

$$C_{L_2}(y) = \sum_t |y_t - \bar{y}|^2 \tag{A.3}$$

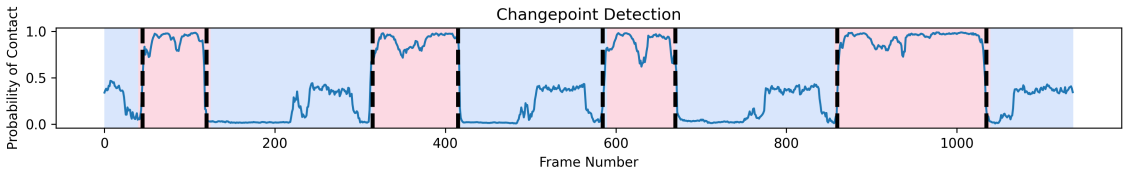


Figure A.4: An example of segmented changepoints using the PELT algorithm. The black dotted lines represent predicted changepoints, and the blue and red regions represent segments between trials and during trials, respectively. The detected changepoints are clearly very close to the true transitions. The blue line, which represents probability of hand-object contact over time, was generated using image classifier 1. The data is the same as in Figure A.1a.

Appendix B: Additional Mathematical Justification

Proof of Equation 2.13

Proof. Let m be an event of interest. Let E and N be mutually exclusive and exhaustive events, i.e., $P(E) + P(N) = 1$. Consider a partition of the sample space into n mutually exclusive and exhaustive events, denoted by ρ_i for $i = 1$ to n. Assume that $P(m|E \cap \rho_i) = P(m|N \cap \rho_i)$. We will show that $P(m|E) = \sum_{i=1}^{n} P(m|\rho_i)P(\rho_i|E)$.

$$
\begin{aligned}
P(m|E) &= \frac{P(m \cap E)}{P(E)} \quad \text{by definition of conditional probability} \\
&= \frac{1}{P(E)} \sum_{i=1}^{n} P(m \cap E \cap \rho_i) \quad \text{by law of total probability} \\
&= \frac{1}{P(E)} \sum_{i=1}^{n} P(m|E \cap \rho_i) P(E \cap \rho_i) \\
\Rightarrow P(m|E) &= \sum_{i=1}^{n} P(m|E \cap \rho_i) P(E|\rho_i) \quad \text{by definition of conditional probability}
\end{aligned}
$$

$$
\text{Where } P(m|\rho_i) = P(m|E \cap \rho_i)P(E|\rho_i) + \underbrace{P(m|N \cap \rho_i)}_{=P(m|E \cap \rho_i)} P(N|\rho_i) \quad \text{by law of total probability}
$$

$$
\begin{aligned}
P(m|\rho_i) &= P(m|E \cap \rho_i)\Big(\underbrace{P(E|\rho_i) + P(N|\rho_i)}_{=1}\Big) \\
\Rightarrow P(m|\rho_i) &= P(m|E \cap \rho_i)
\end{aligned}
$$

$$
\therefore P(m|E) = \sum_{i=1}^{n} P(m|\rho_i)P(\rho_i|E)
$$

□

Appendix C: Additional Temporal Video Segmentation Qualitative Results

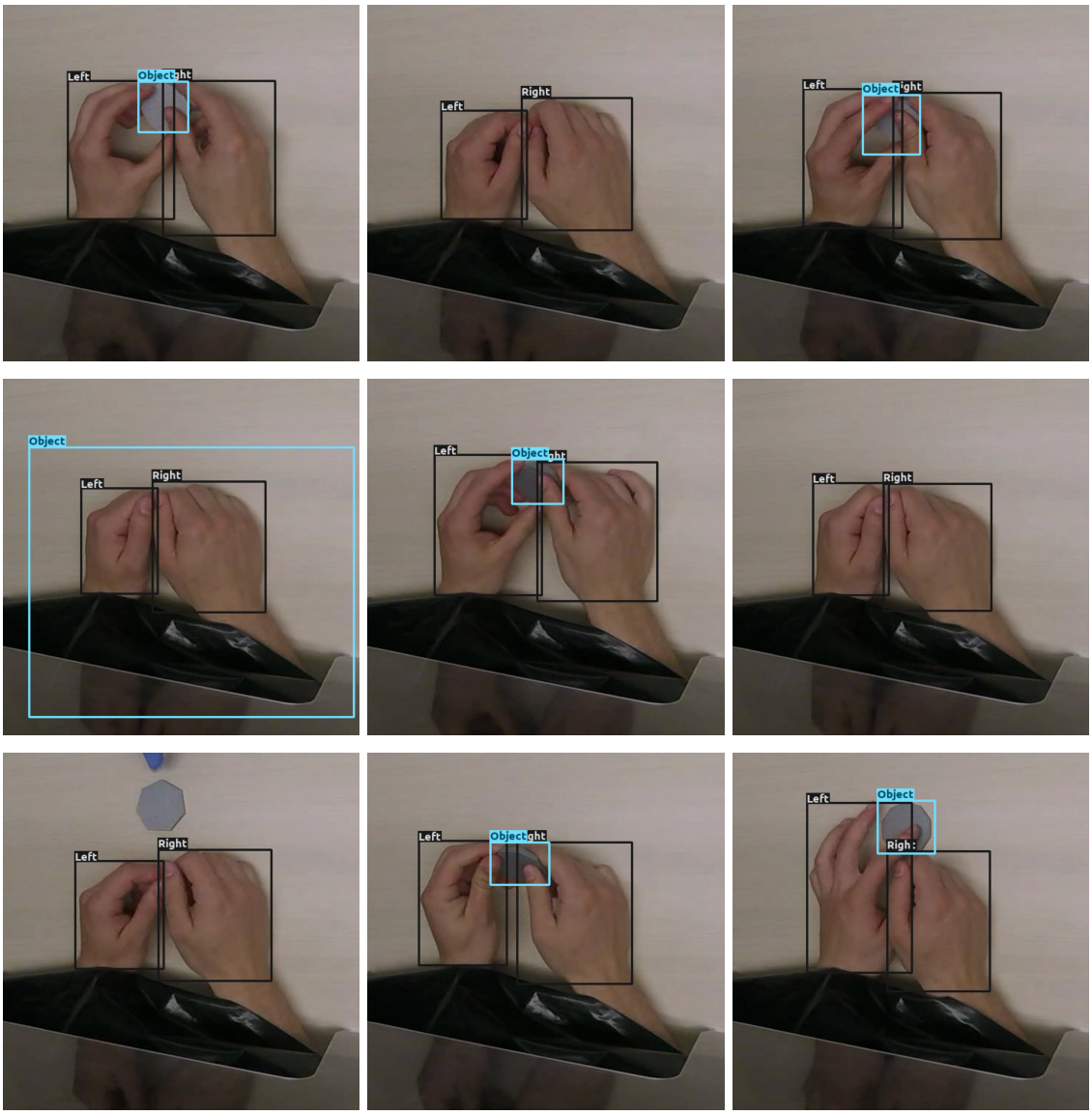


Figure C.1: Frame-Wise Qualitative Results for Volunteer 1

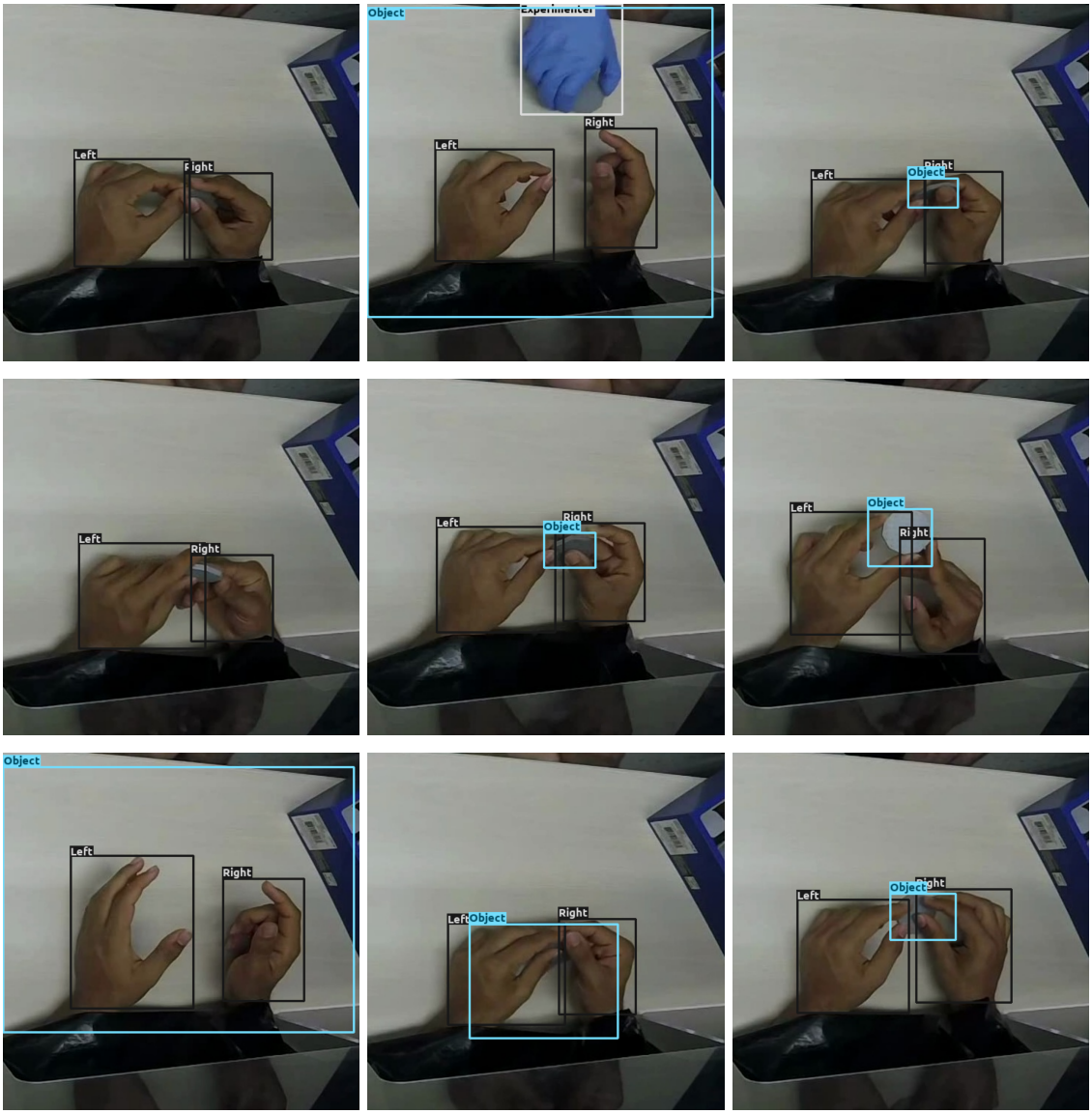


Figure C.2: Frame-Wise Qualitative Results for Volunteer 2

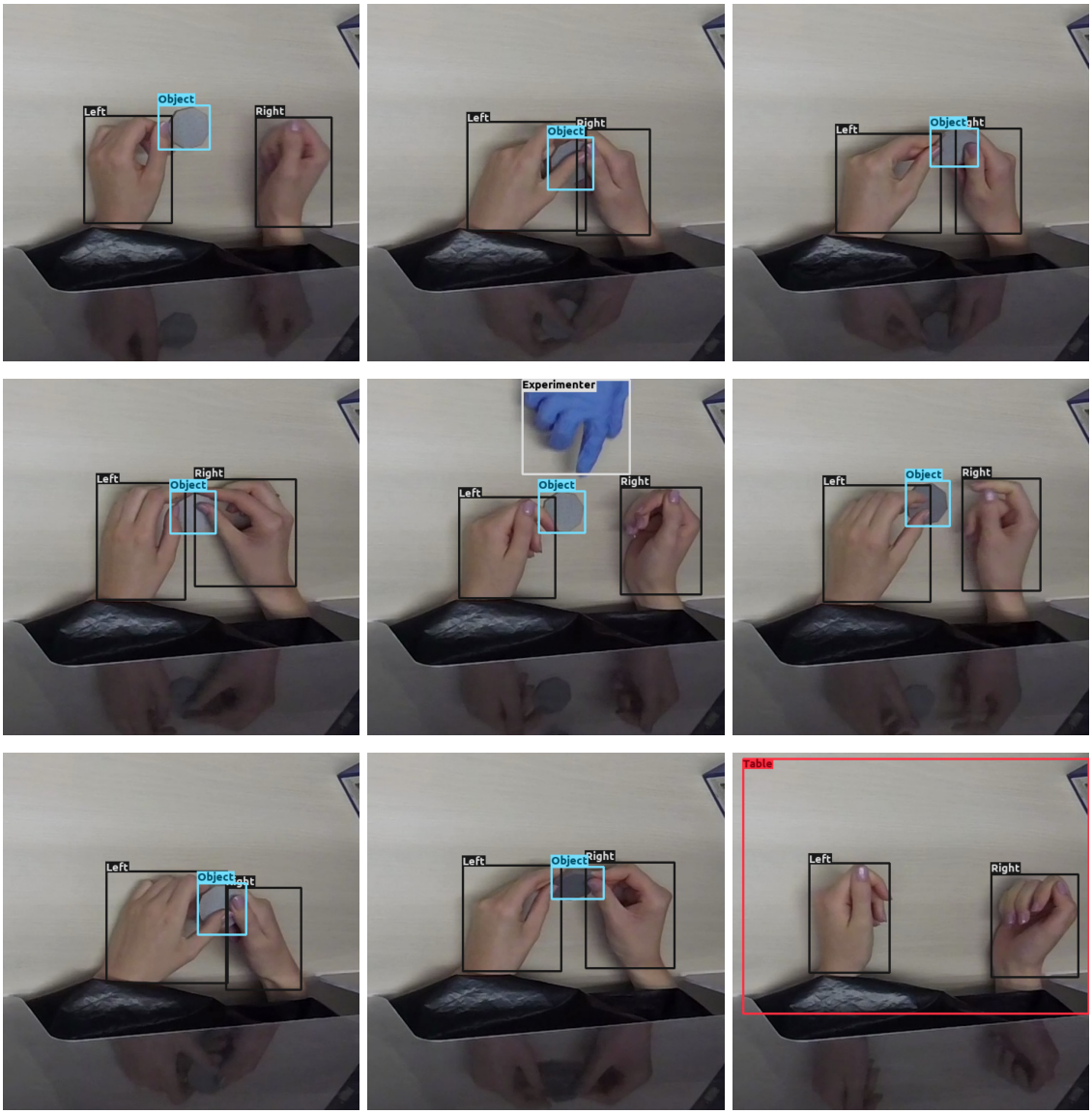


Figure C.3: Frame-Wise Qualitative Results for Volunteer 3

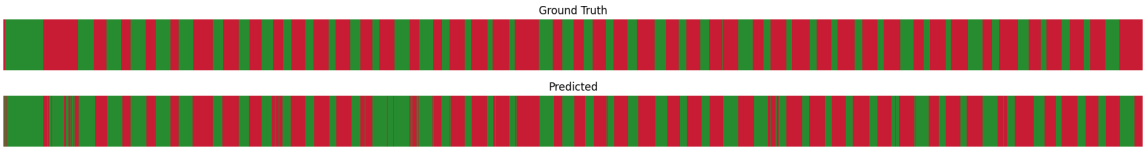


Figure C.4: Whole-Sequence Qualitative Results for Volunteer 1

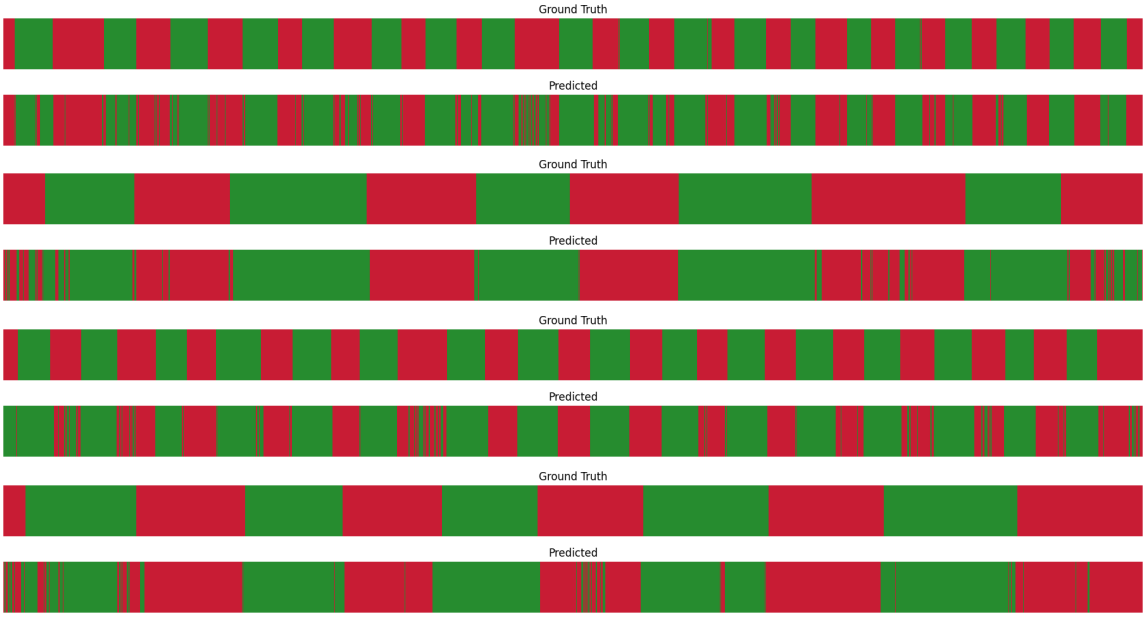


Figure C.5: Whole-Sequence Qualitative Results for Volunteer 2

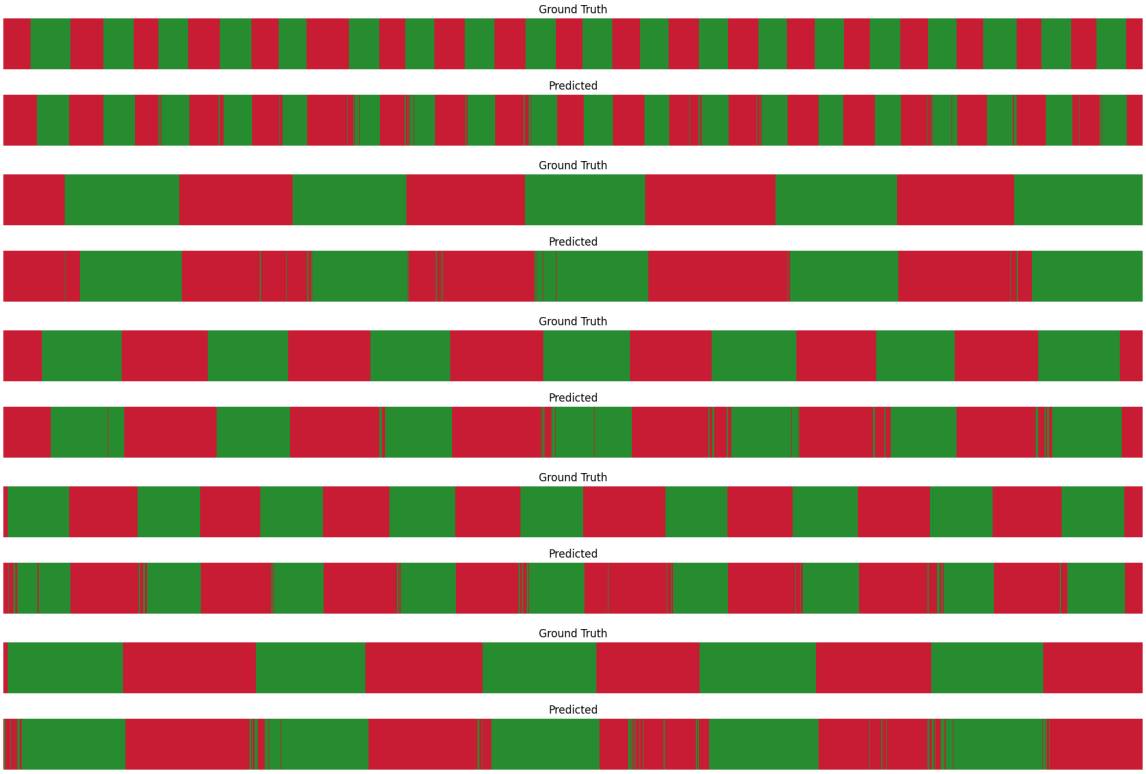


Figure C.6: Whole-Sequence Qualitative Results for Volunteer 3

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257–. https://doi.org/10.1016/S0960-9822(04)00043-0

Amjoud, A. B., & Amrouch, M. (2023). Object detection using deep learning, cnns and vision transformers: A review. *IEEE Access*, *11*, 35479–35516. https://doi.org/10.1109/ACCESS.2023.3266093

Bankieris, K. R., Bejjanki, V. R., & Aslin, R. N. (2017). Sensory cue-combination in the context of newly learned categories. *Scientific reports*, *7*(1), 10890–10. https://doi.org/10.1038/s41598-017-11341-7

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PloS one*, *6*(5), e19812–e19812. https://doi.org/10.1371/journal.pone.0019812

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724–4733.

Chen, Y., Wang, S., & Ge, Y. (2022). A survey on the applications of image classification based on convolution neural network. *2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, 381–384. https://doi.org/10.1109/IPEC54454.2022.9777354

Colombo, M., & Seriès, P. (2012). Bayes in the brain—on bayesian modelling in neuroscience. *The British journal for the philosophy of science*, *63*(3), 697–723. https://doi.org/10.1093/bjps/axr043

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255. 10.1109/CVPR.2009.5206848

Ding, G., Sener, F., & Yao, A. (2024). Temporal action segmentation: An analysis of modern techniques. *IEEE transactions on pattern analysis and machine intelligence*, *46*(2), 1–19. https://doi.org/10.1109/TPAMI.2023.3327284

Farahani, A., Pourshojae, B., Rasheed, K., & Arabnia, H. R. (2021). A concise review of transfer learning. *arXiv.org*. https://doi.org/10.48550/arXiv.2104.02144

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current biology*, *13*(6), 483–488. https://doi.org/10.1016/S0960-9822(03)00133-7

Ishikawa, Y., Kasai, S., Aoki, Y., & Kataoka, H. (2021). Alleviating over-segmentation errors by detecting action boundaries. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2321–2330. https://doi.org/10.1109/WACV48630.2021.00237

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, *107*(500), 1590–1598. https://doi.org/10.1080/01621459.2012.737745

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., & Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC medical imaging*, *22*(1), 69–69. https://doi.org/10.1186/s12880-022-00793-7

Klatzky, R., Lederman, S., & Metzger, V. (1985). Identifying objects by touch: An "expert system". *Perception & Psychophysics*, *37*, 299–302. https://doi.org/10.3758/BF03211351

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lea, C., Flynn, M. D., Vidal, R., Reiter, A., & Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1003–1012. https://doi.org/10.1109/CVPR.2017.113

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature (London)*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Li, Z., Farha, Y. A., & Gall, J. (2021). Temporal action segmentation from timestamp supervision. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8361–8370. https://doi.org/10.1109/CVPR46437.2021.00826

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. In *Lecture notes in computer science* (pp. 21–37). Springer International Publishing. https://doi.org/10.1007/978-3-319-46448-0_2

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Wan-Teh, C., Hua, W., Georg, M., & Grundmann, M. (2019).

Mediapipe: A framework for building perception pipelines. *arXiv.org*. https://doi.org/10.48550/arXiv.1906.08172

O'Reilly, J. X., Jbabdi, S., & Behrens, T. E. J. (2012). How can a bayesian approach inform neuroscience? *The European journal of neuroscience*, *35*(7), 1169–1179. https://doi.org/10.1111/j.1460-9568.2012.08010.x

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-CNN: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

Ribani, R., & Marengoni, M. (2019). A survey of transfer learning for convolutional neural networks. *2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 47–57. https://doi.org/10.1109/SIBGRAPI-T.2019.00010

Shams, L., & Beierholm, U. (2022). Bayesian causal inference: A unifying neuroscience theory. *Neuroscience and biobehavioral reviews*, *137*, 104619–104619. https://doi.org/10.1016/j.neubiorev.2022.104619

Shan, D., Geng, J., Shu, M., & Fouhey, D. F. (2020). Understanding human hands in contact at internet scale. https://doi.org/10.48550/arXiv.2006.06669

Tan, M., & Le, Q. V. (2021). Efficientnetv2: Smaller models and faster training. *arXiv.org*. https://doi.org/10.48550/arXiv.2104.00298

Truong, C., Oudre, L., & Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal processing*, *167*, 107299–. https://doi.org/10.1016/j.sigpro.2019.107299

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, *104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5

Xu, Z., Rawat, Y., Wong, Y., Kankanhalli, M. S., & Shah, M. (2022). Don't pour cereal into coffee: Differentiable temporal logic for temporal action segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 14890–14903, Vol. 35). Curran Associates, Inc. https:

//proceedings.neurips.cc/paper_files/paper/2022/file/5f96a21345c138da929e99871fda138e-Paper-Conference.pdf